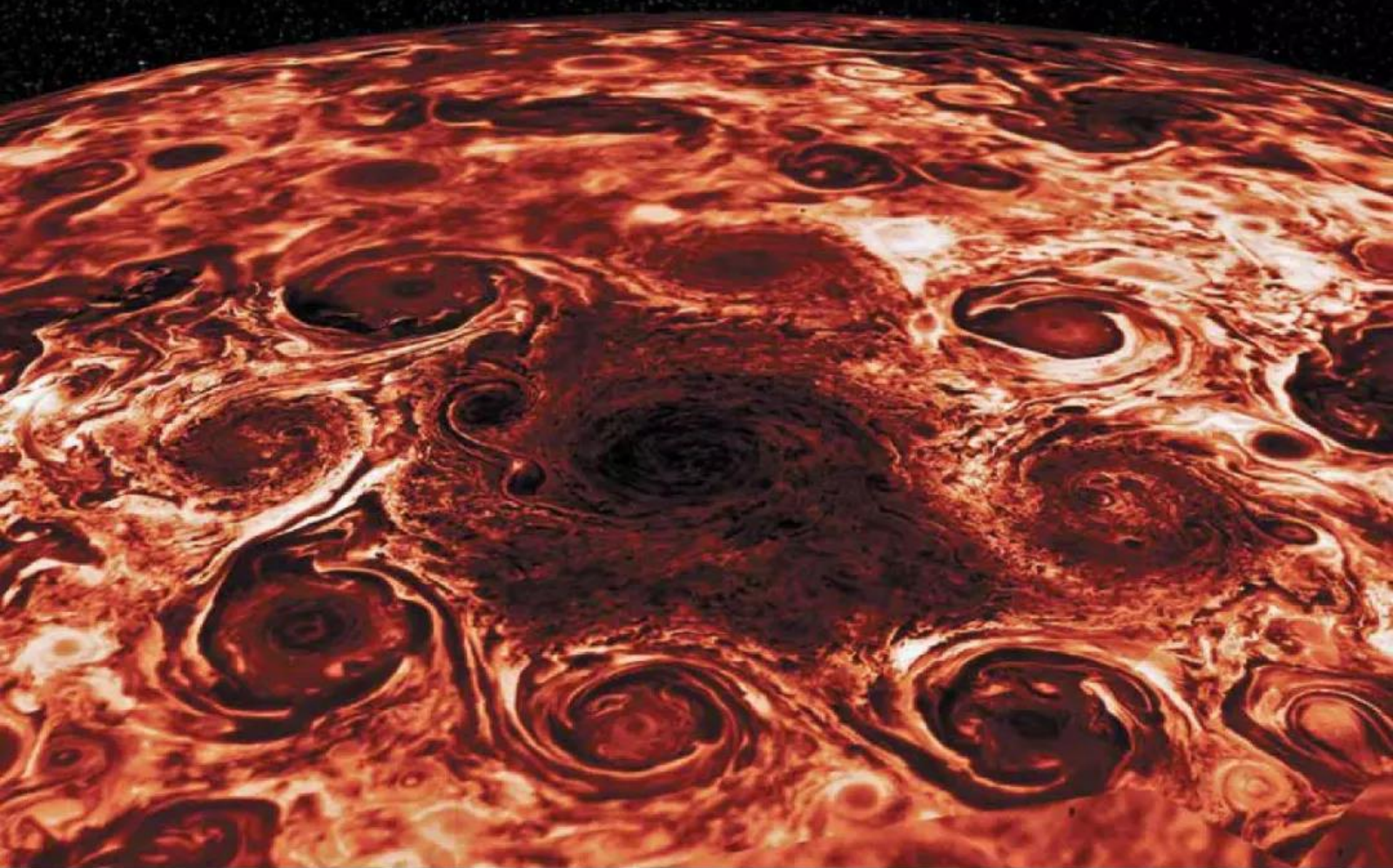


# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

## JUPITER REVISITED

*Juno mission offers a fresh perspective  
on the gas giant* **PAGES 168, 216, 220, 223 & 227**



**CLIMATE CHANGE**

### FLOOD WARNING

*Rising sea levels raise risk  
of extreme events*

**PAGE 156**

**HUMAN BEHAVIOUR**

### PEER REVIEW

*The reputational judgements  
that drive cooperation*

**PAGES 169 & 242**

**EVOLUTION**

### FLYING HIGH

*How mountains shape  
bird diversity*

**PAGES 173 & 246**

**NATURE.COM/NATURE**

8 March 2018

Vol. 555, No. 7695



# THIS WEEK

## EDITORIALS

**PEER REVIEW** Nature journals encourage more scrutiny of code **p.142**



**WORLD VIEW** How golden tickets help science find new jewels **p.143**

**GALAXIES** Distant black hole tore apart star in a flash **p.144**

## A brighter farming future

*A massive, decade-long experiment involving millions of Chinese farmers demonstrates an evidence-based approach to sustainability.*

In 1958, China under Mao Zedong embarked on a nationwide political project to increase agricultural productivity by collectivizing small farms across the country and forcing them to share agricultural tools. It was a disaster and contributed to a famine in which tens of millions died.

Now science has succeeded where ideology failed. A huge, decade-long experiment involving millions of farmers reports its results this week. Writing in *Nature*, scientists in China describe how they identified and passed on evidence-based techniques to make smallholder farming in the country more efficient (Z. Cui *et al.* *Nature* <http://dx.doi.org/10.1038/nature25785>; 2018). No sharing of agricultural tools was required; just the gathering and pooling of scientific data on local conditions and agricultural needs.

Running from 2005 to 2015, the project first assessed how factors including irrigation, plant density and sowing depth affected agricultural productivity. It used the information to guide and spread best practice across several regions: for example, recommending that rice in southern China be sown in 20 holes densely packed in a square metre, rather than the much lower densities farmers were accustomed to using.

The results speak for themselves: maize (corn), rice and wheat output grew by some 11% over that decade, whereas the use of damaging and expensive fertilizers decreased by between 15% and 18%, depending on the crop. Farmers spent less money on their land and earned more from it — and they continue to do so.

The results offer hope in the search for a more sustainable future on a crowded planet. After all, some 2.5 billion smallholders together farm 60% of the world's arable land. Beyond that, the project provides many lessons. First, that a scientific approach can increase agricultural productivity and cut damage to the environment. Second, that such success requires investment in what economists call the intangibles — the creation of networks to spread information and give scientists access to essential data. The scale of the research network created is impressive: 1,200 scientists, 65,000 local officials, 140,000 industry representatives and 21 million farmers across 37.7 million hectares.

Maintaining the people in those networks — in this case, the technicians and bureaucrats in local government offices — is a must. The study shows how these posts can produce benefit, both economic and environmental, far beyond what they cost. Unfortunately, in many countries, such jobs and the networks that depend on them are being cut back, often, paradoxically, in the name of efficiency.

The third lesson is that the same methods could, in principle, be used to boost agricultural efficiency elsewhere. But that will not be easy. China has well-developed regional infrastructure and relatively efficient central control, both of which allowed this project to operate on such a large scale. India and Africa — two regions that could benefit from a similar approach — do not. That makes it difficult, although not impossible, to translate the study and the results beyond China.

Fourth, the programme must be monitored and updated. Its

recommendations were fine-tuned to the needs of farmers in specific regions, but these can change, especially as the climate alters. To consolidate their success, the farmers and scientists involved should continue to adapt the recommended methods.

China must now build on this project. Some 200 million smallholdings are not yet plugged into the information networks set up and so are not applying the recommendations. There is scope for easy wins here.

**“There is a thrill in finding that expectations hold up over so grand a scale.”**

For example, researchers could piggyback on existing but separate networks. One is the Science and Technology Backyard platforms, which operate in 21 provinces and cover a wide range of crops. They bring agricultural scientists to live in villages, and use demonstrations to show farmers better techniques.

Such projects could ensure that farmers continue to learn. They could also be expanded to investigate the best use of other agricultural options, such as pest management and the use of legumes as alternatives to fertilizers.

Perhaps the most important lesson is that better use of existing technology can help to produce more food in a sustainable way. None of the recommendations given to China's farmers would have surprised agronomists. Still, the scientists involved deserve great credit for having the vision and the wherewithal to make the project happen.

There is a thrill in finding that expectations hold up over so grand a scale. And, ultimately, it was that scale that made the difference. It allowed the project to go where even the best smaller studies (and Mao Zedong) could not: persuading often intractable rural farmers to change their practices, and so improve efficiency and productivity. ■

## Learn to tell tales

*Ocean researchers are among those inspired by science fiction to tell diverse tales of the future.*

Among the many things that SpaceX likes to do differently is name its hardware. Last month, chief executive Elon Musk announced that the company's latest droneship (the floating ocean platforms designed to receive reusable rocket launch boosters) will be known as A Shortfall of Gravitas. It will join existing barges Just Read the Instructions and Of Course I Still Love You.

The names will be familiar to readers of the Scottish author Iain M. Banks (who died in 2013) as based on spacecraft from his Culture series of science-fiction novels. And Banks is about to get an even wider audience: tech entrepreneur Jeff Bezos is also a fan, and his firm



Amazon has announced plans to film the first of the Culture books.

The path from science fiction to science fact has been well explored, especially in areas such as space and technology, with inventions from satellites to iPads first imagined in stories. But can the influence go further? What if it is not the concepts described by science fiction that could have the most impact, but the act of storytelling — the creation of scientific narratives — itself?

That's the goal of something called science-fiction prototyping. Developed by Brian David Johnson at computer company Intel a decade or so ago to help the firm's engineers anticipate future demand, the approach takes scientific facts and spins them into the future to explore the societal scenarios that could emerge. Advocates say an emphasis on exploring how humans might react to technological change creates a "focused, tailored and creative way to think about possible futures around a particular issue" (A. Merrie *et al. Futures* **95**, 22–32; 2018). It differs from other forms of scenario planning, they argue, because the emphasis is placed as much on the narrative used to explore the results as on the results themselves, and because the goal is not to reach a predetermined outcome. The method has been used by researchers at the University of Essex, UK, and King Abdulaziz University in Jeddah, Saudi Arabia, to create and test a virtual-reality-based distance-learning tool originally imagined for the year 2048 that they call the BReal Lab ([go.nature.com/2fhz9za](http://go.nature.com/2fhz9za)).

Sustainability scientist Andrew Merrie at Stockholm University and his colleagues have taken this principle and applied it to a topical environmental concern: the fate of the world's oceans. The project paints four scenarios for 2050–70, each of which builds on current trends in oceans governance and the fishing industry, as well as ongoing development of marine science and technology. More-uncertain outcomes — the possible collapse of ice sheets and the formation of deep-sea dead zones as a result of onshore pollution — play out

differently for better and worse.

One scenario, called Oceans Back from the Brink, describes a public talk given in 2070 about how an artificial-intelligence system released all forms of confidential data, which prompted the collapse of existing corporate structures and renewed conservation efforts. Another, Rime of the Last Fisherman — Dispatches from a Dying Ocean, imagines a less-than happy ending, with decaying oceans, a geoengineering experiment gone badly wrong and onshore disaster.

**“Narrative has an important role in the communication of science.”**

The paper in *Futures* is accompanied by striking illustrations on the project's website ([go.nature.com/2orkrux](http://go.nature.com/2orkrux)).

Narrative has an important role in the communication of science, but can it also help in the pursuit of research? Purists may baulk, but stories already feature heavily, from the promised potential of work pitched in grant applications to the case studies of impact that funders increasingly ask for when projects finish. Climate-change science has long relied on emissions scenarios that diverge according to how future societies might behave. These rely not on extrapolation of current trends, but on imagined differences in, for example, whether nations come to cooperate or opt to pursue their own agendas. And climate-change policies are being planned on the basis of stories of future technology — carbon capture and negative-emissions equipment included — that many argue are pure fiction and will never materialize.

Some of the scenarios painted — in both the fictional tales of the future ocean and the high-emissions scenarios of climate modellers — are something that society, scientists included, should be desperate to avoid. To do so, data and evidence remain the priority. But in a world where both are so easily trumped by a seductive (and false) counter-narrative, perhaps more researchers should also learn to tell tales as they look ahead. ■

## Code check

*Researchers who rely on bespoke software are encouraged to submit the programs for scrutiny.*

Computer code written by scientists forms the basis of an increasing number of studies across many fields — and an increasing number of papers that report the results. So, more papers should include these executable algorithms in the peer-review process. From this week, Nature journal editors handling papers in which code is central to the main claims or is the main novelty of the work will, on a case-by-case basis, ask reviewers to check how well the code works, and report back.

The move builds on growing demand in recent years for authors to publish the details of bespoke software used to process and analyse data. And it aims to make studies that use such code more reliable. Computational science — like other disciplines — is grappling with reproducibility problems, partly because researchers find it difficult to reproduce results based on custom-built algorithms or software.

This policy is the latest stage in the evolution of our editorial processes, which aims to keep up with technological change across the research community. All Nature journals, for example, already require that authors make materials, data, code and associated protocols promptly available to readers on request, without undue qualifications. In 2014, the Nature journals adopted a “code availability” policy to ensure that all studies using custom code deemed central to the conclusions include a statement indicating whether and how the code can be accessed, and explain any restrictions to access.

Some journals have for years gone a step further and ensured that the new code or software is checked by peer reviewers and published

along with the paper. When relevant, *Nature Methods*, *Nature Biotechnology* and, most recently, journals including *Nature* and *Nature Neuroscience* encourage authors to provide the source code, installation guide and a sample data set, and to make this code available to reviewers for checking.

To assist authors, reviewers and editors, we have updated our guidelines to authors ([go.nature.com/2d2i80d](http://go.nature.com/2d2i80d)) and have developed a code and software submission checklist ([go.nature.com/2h9ouaj](http://go.nature.com/2h9ouaj)) to help authors compile and present code for peer review. We also strongly encourage researchers to take advantage of repositories such as GitHub, which allow code to be shared for submission and publication.

According to the guidelines, authors must disclose any restrictions on a program's accessibility when they submit a paper. *Nature* understands that in some cases — such as commercial applications — authors may not be able to make all details fully available. Together, editors and reviewers will decide how the code or mathematical algorithm must be presented and released to allow the paper to be published.

Occasionally, other exceptions will be made — for example, when custom code or software needs supercomputers, specialized hardware or very lengthy running times that make it unfeasible for reviewers to run the necessary checks. We also recognize that preparing code in a form that is useful to others, or sharing it, is still not common in some areas of science.

Nevertheless, we expect that most authors and reviewers will see value in the practice. Last year, *Nature Methods* and *Nature Biotechnology* between them published 47 articles that hinged on new code or software. Of these, approximately 85% included the source code for review.

As with other scientific fields, the impact of computational tools is determined by their uptake. Open implementation increases the likelihood that other researchers can use and build on techniques. So, although many researchers already embrace the idea of releasing their code on publication, we hope this initiative will encourage more to do so. ■





## Fund ideas, not pedigree, to find fresh insight

*Anonymous applications free scientists to make bold proposals, and ‘golden tickets’ free reviewers to bet on them, says Thomas Sinkjær.*

About five years ago, when I was director of the Danish National Research Foundation in Copenhagen, I held focus groups to ask postdocs and early-career researchers how funders might further their work. Members of the board and I spoke with more than 400 young scientists and kept hearing the same depressing refrain: many were writing grants not for work they really wanted to do, but for projects they thought could get funded. Often, they were not even bringing their best ideas to the table.

And why would they? Grant review tends to be biased against innovation; researchers' best shot at funding is proposing the same sort of work that they have already proved they can do. Although there is some evidence to suggest that peer review can distinguish solid research from poor research, it is not clear that it can identify the very best — especially as falling funding rates demand that reviewers make finer and finer distinctions when selecting which projects to support.

One way to improve the situation is for funders to try different schemes and share their experiences. The Villum Fonden is the largest philanthropic foundation in Denmark for the support of technical and natural-science research. Such foundations have more leeway than organizations funded by taxpayers to experiment with different ways of selecting which research to finance.

Two years ago, when I was director of science at the foundation, we set up a project that we hoped would support innovative ideas by evaluating applications in an unusual way. Assessment of research proposals would be blinded and based on a three-page description. Evaluators would have no information on the applicant's background or publishing record. By coincidence, I learned that the Volkswagen Foundation in Hanover, Germany, was running a similar scheme; we both hoped to gather evidence on how grant review worked. Each foundation had, independently, dubbed its new scheme 'Experiment'.

In January 2017, the Villum Experiment called for "science so risky that applicants would not normally consider putting forward the project for funding". We committed about 15% of our annual funds to this sort of research. We recruited evaluators whom we thought (by reputation) would be particularly able to judge risky ideas — for example, people we knew to have discussed new ways of funding research. They ranked each application they read. Each reviewer was also given one 'golden ticket' — a right to fund an application, no matter what their fellow reviewers said.

Funding rates at both foundations were just over 10% of the applications submitted for this call. Recipients included both postdocs and department heads, and about one-third of successful applicants were under the age of 40. So far, the Villum Foundation has awarded 39 grants of up to two years each, and the Volkswagen scheme 96 of up to 18 months; overall, each grant is worth from about €120,000

(US\$148,000) to €250,000. In the Villum model, 31% were funded on the basis of golden tickets. Although all golden-ticket grants scored better than most others in this call, about half would not have been funded if based on cumulative scores from all reviewers. In the Volkswagen scheme, 11% were golden tickets, none of which would have been funded otherwise.

In a survey, about half of the recipients said that had the call for unorthodox ideas not been anonymous, they would not have proposed their winning idea — they didn't think they had a shot if judged on their publishing track records. Reviewers said that they liked evaluating ideas without knowing the applicant's past performance.

There are wrinkles to iron out. Some reviewers are concerned that if junior researchers' risky ideas don't work out, promising scholars will

have missed a chance to pursue more-conservative projects. Others warn that recipients might not be qualified to carry out their plans. It is too soon to know, and we want to learn more. The second round of applications closes on 21 March.

Meanwhile, we want to work out how to bring in more ideas. We asked applicants what might have kept colleagues from applying. Answers included discomfort with risky projects; concerns that funding decisions would be haphazard; short-duration and limited funds; the inability to simply reuse another application; and a perception that ideas were either not good or risky enough. The numbers are too small to be certain, but there are signs that men are more likely to get funded. Both foundations plan to tweak how applicants and reviewers are recruited — for example, using ungendered text in the call for proposals — and

will continue to monitor diversity.

What interests me most about the experiment is the prospect of better understanding peer review to improve the process. The Bill & Melinda Gates Foundation uses blind review for awards in its Grand Challenges Explorations programme, and New Zealand's Health Research Council uses a random-number generator to prioritize 'Explorer' grant proposals that have fulfilled certain criteria. The global RAND Corporation and an international panel convened by the Canadian Institutes of Health Research have compiled an overview of review approaches and the — limited — empirical evidence for them.

To paraphrase Winston Churchill, grant peer review might be the worst system, except all the others. Given the massive resources dedicated to it, we need a better evidence base to guide its evolution. ■

**Thomas Sinkjær** is professor at Aalborg University and senior vice-president for grants and prizes at the Lundbeck Foundation in Copenhagen.  
e-mail: [ts@lundbeckfonden.com](mailto:ts@lundbeckfonden.com)

HALF OF THE  
RECIPIENTS SAID HAD  
THE CALL NOT BEEN  
**ANONYMOUS,**  
THEY WOULD  
NOT HAVE  
PROPOSED THEIR  
**WINNING IDEA.**



# SEVEN DAYS

The news in brief

## PEOPLE

### Leadership change

The head of the National Natural Science Foundation of China has stepped down, the government announced on 27 February. Yang Wei, a materials scientist, had led the funding agency since 2013 and had become a crusader for research integrity in China. He will be replaced by chemical engineer Li Jinghai. Li is a former vice-president at the Chinese Academy of Sciences and a current vice-president at the International Council for Science. NSFC grants are the main source of funding for researchers in China. In 2017, 10.7 billion yuan (US\$1.69 billion) from the agency supported more than 18,000 research projects. Yang told *Nature* he plans to return to academic research in soft-matter physics.

### Iranian scholar

Ahmadreza Djalali, a disaster-medicine researcher sentenced to death in Iran for spying, protested his innocence in a letter to the country's president, Hassan Rouhani, on 4 March. Djalali, an Iranian who lived in Sweden, was arrested in Tehran in 2016 and convicted for espionage last October. In February, he filed a libel complaint against the spokesperson of the Iranian foreign affairs ministry over statements that linked him to the deaths of two Iranian nuclear scientists. In the letter, Djalali says that he has been subjected to an unjust judicial process and has never confessed to being a spy or been involved in the assassination of Iranian scholars. He has twice unsuccessfully appealed against his death sentence, but he can still appeal to other branches of Iran's Supreme

Court. Sweden has granted Djalali citizenship as a sign of support.

### German minister

In a surprise move, German Chancellor Angela Merkel has chosen little-known lawmaker Anja Karliczek as the country's next minister of education and research. Karliczek, a Christian Democrat who has been a member of the German parliament since 2013, has a background in banking and business but little experience in science and education policy. She succeeds Christian Democrat Johanna Wanka, who had served in the post since 2013. The selection must be approved by Germany's incoming government. On 4 March, the Social Democrats

voted in favour of forming a new government coalition with Merkel's Democratic Union, ending a five-month political impasse.

## UNIVERSITIES

### Harassment report

The University of California, Berkeley, mishandled reports of sexual misconduct by staff and students, the US Department of Education said on 28 February after a four-year investigation. The probe began after 31 students filed a complaint in 2014, alleging that the university had violated federal laws against sex discrimination. The government found that the institution had not always resolved complaints in a

timely manner or provided opportunities for formal investigations. The university will amend its policies, re-examine eight cases of alleged sexual misconduct and submit to government monitoring for two years. "We remain committed to doing more to improve our processes and ensure a safe and supportive environment," university chancellor Carol Christ said in a statement.

## FUNDING

### Budget boost

Hong Kong will invest more than 50 billion Hong Kong dollars (HK\$; US\$6.4 billion) in science, technology and innovation. This year's budget was announced on



RACHAEL HERMAN, LOUISIANA STATE UNIV./STONY BROOK UNIV.

## Surprise Adélie penguin population

Using aerial drones together with field expeditions, scientists have confirmed that there are more than 1.5 million Adélie penguins (*Pygoscelis adeliae*) on the Danger Islands, a remote island chain off the western coast of Antarctica. The penguin group, one of the world's largest, has not suffered the population declines seen in other areas of the Antarctic

Peninsula, the researchers reported on 2 March (A. Borowicz *et al. Sci. Rep.* <http://doi.org/ck5j>; 2018). Scientists found the colony by combining drone surveys and field expeditions to validate satellite images. The animals might have gone undetected for decades because thick sea ice surrounding the islands made them difficult to reach, the scientists say.



28 February and builds on the HK\$10 billion invested in these areas last year. The government will allocate HK\$20 billion to the upcoming Hong Kong–Shenzhen Innovation and Technology Park, to promote ventures between Chinese and international companies. Another HK\$10 billion will fund new research centres specializing in artificial intelligence, robotics and health care. Technology companies can tap into a HK\$500-million-grant offering subsidies for hiring postgraduate students.

## BUSINESS

## Singapore deal

Chinese tech giant Alibaba will open its first research centre outside China at Singapore's Nanyang Technological University (NTU). The company announced on 28 February that the Alibaba–NTU Singapore Joint Research Institute will initially involve 50 researchers from both organizations, and it will focus on machine learning, cloud computing and natural-language-processing capabilities. Alibaba's backing for the five-year partnership is estimated to be worth more than 42.5 million Singapore dollars (US\$32 million), and is NTU's largest investment from a company so far. The funding

comes from Alibaba's US\$15-billion global research and development programme.

## SPACE

## Satellite launch

NASA launched the United States' latest weather satellite on 1 March. The Geostationary Operational Environmental Satellite-17 (GOES-17) will operate at 35,800 kilometres above the equatorial Pacific, and it is the most recent in a series of geostationary satellites operated by the National Oceanic and Atmospheric Administration. The craft will monitor phenomena including wildfires, snow cover, clouds and storms. Together with the identical GOES-16, which is already in position over the Atlantic, the satellites will provide weather forecasters with a view extending from the west coast of Africa to the United States and New Zealand. See page 154 for more.

## Red rover

NASA has tested an alternative way for its Mars Curiosity rover (**pictured**) to collect samples on the red planet. The rover's drill stopped working in December 2016 because it could no longer be extended and retracted properly. Curiosity will now push its entire robotic arm forward to



press the drill bit against the rock, NASA announced on 28 February. Since landing in 2012, the rover has drilled and sampled only 15 times. It is currently exploring a region of Mars known as Vera Rubin Ridge.

## EVENTS

## Genetic predictions

The US National Institutes of Health will launch a two-year pilot project to study the effects of gene variants on various human traits. The Genomic Ascertainment Cohort project, announced on 1 March, will use genomic data from 10,000 individuals to make predictions about traits such as height and blood type. Investigators will then re-examine study participants

who have consented to be contacted again, to test those predictions. The project will be carried out in conjunction with Inova, a health-care organization headquartered in Falls Church, Virginia.

## Italian boycott

Italy's National Institute of Health boycotted a national biology conference on 2 March over concerns that anti-vaccine ideas would be promoted. At the meeting — organized by the National Order of Biologists (ONB), a professional organization — Nobel-prizewinning virologist Luc Montagnier and ONB president Vincenzo D'Anna questioned the safety of some vaccines and criticized compulsory immunization programmes. The conference programme sparked uproar in Italy's scientific community when it was announced in January. Academics and scientific societies, including the Italian Society of Microbiology and the National Board of Physicians and Surgeons, had urged the ONB to revise the meeting agenda.

## Integrity office

The CNRS, France's national research agency, will create a research-integrity office. Agency chief Antoine Petit said that the lack of such a body at the CNRS, which has 32,000 staff members and is Europe's largest basic-research agency, was significant. He said that he was keen to uphold due process in misconduct investigations, and that any disciplinary actions should be serious but proportional. Petit also emphasized that the agency must create a culture that discourages misconduct. A working group chaired by Olivier Le Gall, head of the French Office for Scientific Integrity, will propose details of the body's structure within three months. Petit announced the initiative on 1 March.

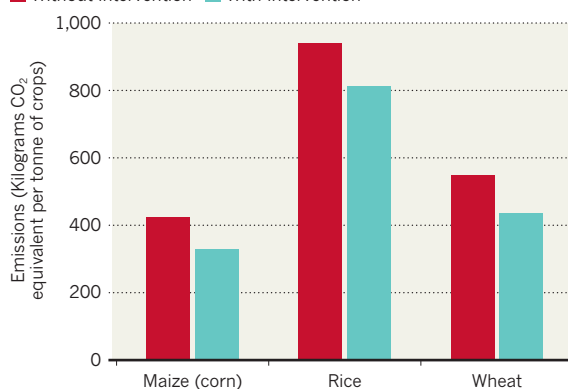
## TREND WATCH

A study in China shows that farming interventions such as applying fertilizer at specific times in the growing cycle can decrease the amount of fertilizer that must be used, increase crop yields and reduce greenhouse-gas emissions. The large study, carried out in 2005–15 and published on 7 March in *Nature*, aimed to improve the sustainability and efficiency of nearly 21 million smallholder farms. Small farms in China use nitrogen fertilizers at four times the global average. See page 141.

## REDUCING FARM EMISSIONS

Managing fertilizer use on small farms in China reduced greenhouse-gas emissions.

■ Without intervention ■ With intervention



# NEWS IN FOCUS

**POLICY** EU pesticide review could lead to ban on neonicotinoids **p.150**

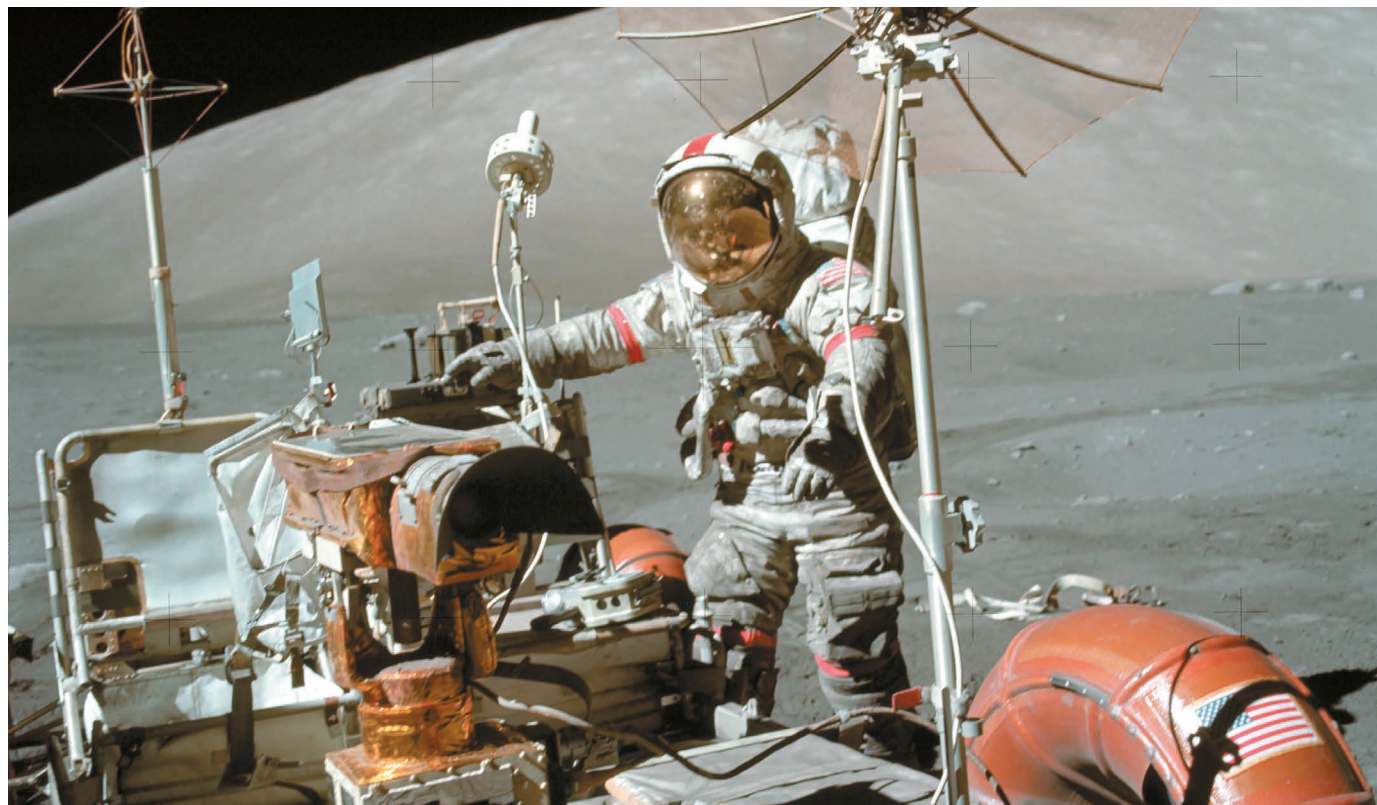
**MATERIALS** Graphene sandwich with a twist acts as a superconductor **p.151**

**POLLUTION** China tests giant air-filter system to clean up urban smog **p.152**



**CLIMATE CHANGE** Once rare, extreme floods are becoming the norm **p.156**

NASA



NASA astronaut Eugene Cernan led the Apollo 17 mission to the Moon in 1972.

## PLANETARY SCIENCE

# US scientists plot return to the Moon's surface

*Lunar researchers seize on Trump administration's political interest in exploration.*

BY ALEXANDRA WITZE

When Apollo astronaut Gene Cernan stepped off the Moon in December 1972, it marked the end of US researchers' access to the lunar surface. Since then, no US mission has touched down there to collect scientific data.

That could soon change. In December, US President Donald Trump ordered NASA to send astronauts back to the Moon. On 12

February, he proposed a 2019 budget that would allow the agency to begin planning a US\$200-million lunar exploration programme. In the weeks since, NASA officials have started sketching out how that effort might unfold — from a series of small commercial landers, to larger NASA landers, to a multinational space station near the Moon that could serve as a base for robots and astronauts travelling to the lunar surface.

For US Moon researchers, Trump's plan

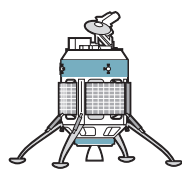
is the first chance for an extended research programme since President Barack Obama cancelled exploration plans in 2010. "It's an exciting time to be a lunar scientist," says Ryan Watkins, a Moon expert at the Planetary Science Institute who works in St Louis, Missouri.

Congress has yet to approve either the president's budget request or his nominee to lead NASA, Representative James Bridenstine (Republican, Oklahoma). But for now, the agency's acting administrator is moving ahead ►



## BACK TO THE MOON

NASA may send landers to the lunar surface for the first time since the 1970s, as part of a renewed exploration programme. Possible spacecraft include:

**MX-1 LANDER****Builder:**

Moon Express of Cape Canaveral, Florida

**Destination:**

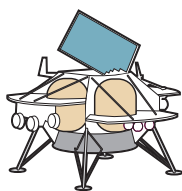
Anywhere on the lunar surface

**Capabilities:**

Carry small payloads selected through a NASA competition

**Time frame:**

As early as 2019

**PEREGRINE LANDER****Builder:**

Astrobot of Pittsburgh, Pennsylvania

**Destination:**

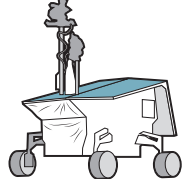
Anywhere on the lunar surface

**Capabilities:**

Carry small payloads selected through a NASA competition

**Time frame:**

As early as 2019

**LUNAR ROVER****Builder:**

NASA

**Destination:**

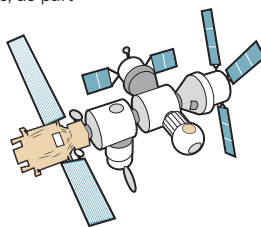
Anywhere on the lunar surface

**Capabilities:**

Study or collect rock samples from different geological formations

**Time frame:**

Early 2020s

**LUNAR ORBITAL PLATFORM-GATEWAY****Builder:**

International space agencies

**Destination:**

Lunar orbit

**Capabilities:**

Provide laboratory and storage space for samples, a refuelling point for surface operations and a communications relay

**Time frame:**

Mid-2020s

▶ with the lunar push (see 'Back to the Moon').

Over the past decade, NASA has sent the Lunar Reconnaissance Orbiter to map the Moon; the Lunar Crater Observation and Sensing Satellite to crash land near the south pole in search of water; the Gravity Recovery and Interior Laboratory to plumb the Moon's gravity field; and the Lunar Atmosphere and Dust Environment Explorer (LADEE) to study its tenuous outer atmosphere.

These and other missions have opened new areas of research, says Dana Hurley, a planetary scientist at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland. "Our understanding has evolved so much in the last decade," she says.

Take LADEE, which detected traces of water in the Moon's atmosphere that were probably carried there by meteorites. Researchers need more detailed data to better understand how water moves around on the lunar surface and

into the atmosphere. "We didn't even know to ask those questions before," Hurley says.

She and other US scientists, in a collaboration known as the Lunar Exploration Analysis Group, have been studying how future missions might answer key science questions. Getting better dates for impact craters on the Moon, for instance, could help establish whether the Solar System experienced a cataclysmic meteorite bombardment 4 billion years ago.

"To take the next really big leaps in lunar science is going to take landing on the ground and getting at it with instruments in a way very similar to what we've done for Mars," says Barbara Cohen, a planetary scientist at NASA's Goddard Space Flight Center in Greenbelt, Maryland, who has developed methods for dating planetary samples on the surface of other worlds (B. A. Cohen *et al.* *Geostand. Geoanal. Res.* **38**, 421–439; 2014). "We have a lot of pent-up demand."

For the first time, NASA might use commercial landers to reach the lunar surface. Companies such as Moon Express of Cape Canaveral, Florida, and Astrobot of Pittsburgh, Pennsylvania, have been developing small landers. Neither they nor their competitors were able to claim the \$30-million Google Lunar XPRIZE, a private effort to put landers and rovers on the Moon by the end of this month. Still, many expect NASA to call in the coming months for proposals that rely on small commercial landers.

"This is the right place. This is the right time," NASA's Sarah Noble told a planetary-science advisory committee on 21 February. "We are really poised to take advantage of this next era of lunar exploration and the opportunities these commercial companies are going to open up"

The first lander missions would probably be short-lived trips to sites on the Moon's near side. But scientists could piggyback on those trips to study topics such as the plasma environment around the lunar poles, or to begin establishing a network of geophysical landers that would listen for moonquakes. By the mid- to late 2020s, NASA might be able to bring samples back to Earth via the space station orbiting the Moon.

Other nations will grab the lunar limelight much sooner. India is slated to launch its Chandrayaan-2 rover later this year to explore near the Moon's south pole. And China is planning to send its Change-4 rover to the lunar far side — a first for any space agency — by the end of 2018.

NASA's challenge will be to keep its latest initiative from falling by the wayside, as did its last big lunar programme — which ran from 2004 to 2010. "I'm excited about the lunar exploration campaign, but concerned we're not making enough investments to get to the surface," says David Kring, a planetary scientist at the Lunar and Planetary Institute in Houston, Texas. He notes that Trump directed astronauts to the Moon; robotic landers do not achieve that goal, no matter how much data they collect. ■

SOURCE: NASA

## POLICY

# EU pesticide review could lead to ban

*Major assessment concludes that neonicotinoids harm bees.*

BY DECLAN BUTLER

In a long-awaited assessment, the European Union's food-safety agency has concluded that three controversial neonicotinoid insecticides pose a high risk to wild bees and honeybees. The findings by

the European Food Safety Authority (EFSA) in Parma, Italy, increase the chances that the EU will soon move to ban all uses of the insecticides on outdoor crops.

In 2013, the EU prohibited use of the three chemicals on crops attractive to bees — such as sunflowers, oilseed rape and maize (corn)

— after an EFSA assessment raised concern about the insecticides' effects. Since then, researchers have amassed more evidence of harm to bees, and the European Commission last year proposed banning all outdoor uses, while still allowing the pesticides in greenhouses. The latest EFSA assessment strengthens the scientific basis for the proposal, says Anca Păduraru, a European Commission spokesperson for public health and food safety. EU member states could vote on the issue as soon as 22 March.

Neonicotinoids (often abbreviated to neonics) are highly toxic to insects, causing paralysis and death by interfering with the central nervous system. Unlike pesticides that remain on plant surfaces, neonicotinoids are taken up throughout the plant — in the roots, stems, leaves, flowers, pollen and nectar.

The EFSA assessment covered the three neonicotinoids of greatest concern to bee health — clothianidin, imidacloprid and thiamethoxam. The agency considered more than 1,500 studies, including all the relevant published scientific literature, together with data from academia, chemical companies, national authorities, non-governmental organizations (NGOs) and beekeepers' and farmers' associations. The assessment found that each of the three chemicals posed at least one type of high risk to bees in all outdoor uses.

The agency found that foraging bees are exposed to harmful levels of pesticide residues in pollen and nectar in treated fields and nearby contaminated areas, as well as in dust created when treated seeds are planted. It also concluded, on the basis of more limited evidence, that neonicotinoids can sometimes persist and accumulate in the soil, and so can affect generations of planted crops and the bees that forage on them.

"EFSA's advice is often criticized by interested parties such as NGOs and companies, but this is a good demonstration of how EFSA gives scientifically sound and impartial advice," says José Tarazona, head of the agency's pesticides unit.

A spokesperson for the global biotechnology firm Syngenta, which produces neonicotinoids, says that EFSA's conclusions are overly conservative. "When regulators make decisions



Honeybees can be exposed to harmful levels of neonicotinoids in pollen, according to an EU review.

about crop-protection products, what should matter is science, data and that the processes in place are respected and that the public interest is served," the spokesperson says. "Any further restrictions based on this report would be ill-conceived."

EU member states were scheduled to vote on the proposal to outlaw outdoor uses on 13 December, but postponed the vote partly

because many wanted to wait until EFSA completed its evaluation.

Member states plan to discuss the EFSA assessment at a meeting of the commission's Standing Committee on Plants, Animals, Food and Feed sometime in March, says Păduraru. "The protection of bees is an important issue for the commission since it concerns biodiversity, food production and the environment." ■

## MATERIALS SCIENCE

# Graphene is a surprise superconductor

*Misaligned sheets of the carbon material can conduct electricity without resistance.*

BY ELIZABETH GIBNEY

A sandwich of two graphene layers can conduct electrons without resistance if they are twisted at a 'magic angle', physicists have discovered. The finding could prove to be a significant step in the decades-long search for room-temperature superconductors.

Most superconductors work only at temperatures close to absolute zero. Even 'high-temperature' superconductors conduct electricity without resistance only at temperatures of up to around  $-140^{\circ}\text{C}$ . A material that displayed the property at room temperature — eliminating the need for expensive cooling — could revolutionize energy transmission, medical scanners and transport.

Physicists now report that arranging two

layers of atom-thick graphene so that the pattern of their carbon atoms is offset by an angle of  $1.1^{\circ}$  makes the material a superconductor. And although the system still needs to be cooled to 1.7 degrees above absolute zero, the results suggest that it might conduct electricity much like known high-temperature superconductors — and that is exciting physicists. The findings were published in two *Nature* papers<sup>1,2</sup> on 5 March.

If confirmed, this discovery would be "very important" to the understanding of high-temperature superconductivity, says Elena Bascones, a physicist at the Institute of Materials Science of Madrid.

Superconductors come broadly in two types: conventional, in which the activity can be explained by the mainstream theory of

superconductivity, and unconventional, where it can't. The latest studies suggest that graphene's superconducting behaviour is unconventional — and has parallels with that of other unconventional superconductors, called cuprates. These complex copper oxides have been known to conduct electricity at up to 133 degrees above absolute zero. And although physicists have focused on cuprates for three decades in their search for room-temperature superconductors, the underlying mechanism has baffled them.

In contrast to cuprates, the stacked graphene system is relatively simple and the material is well-understood. "The stunning implication is that cuprate superconductivity was something simple all along," says Robert Laughlin, a physicist at Stanford University in California.

Graphene already has impressive ►



► properties: its sheets are stronger than steel and conduct electricity better than copper. It has shown superconductivity before<sup>3</sup>, but that occurred in contact with other materials, and the behaviour could be explained by conventional superconductivity.

Physicist Pablo Jarillo-Herrero at the Massachusetts Institute of Technology (MIT) in Cambridge and his team weren't looking for superconductivity when they set up their experiment. Instead, they were exploring how the orientation dubbed the magic angle might affect graphene. Theorists have predicted that offsetting the atoms between layers of 2D materials at this particular angle might induce the electrons that zip through the sheets to interact in interesting ways — although they didn't know exactly how.

The team immediately saw unexpected behaviour in its set-up. First, measurements suggested that the construction had become a Mott insulator<sup>2</sup>. These materials have all the ingredients to conduct electrons, but

interactions between the particles stop them from flowing. Next, the researchers applied an electric field to feed a few extra charge carriers into the system, and it became a superconductor<sup>1</sup>. The existence of an insulating state so close to superconductivity is a hallmark of cuprates

**“These new experiments give cause for cautious celebration.”**

and other unconventional superconductors. Although graphene shows superconductivity at a very low temperature, it does so with just one-ten-thousandth of the electron density of conventional superconductors that gain the ability at the same temperature. In conventional superconductors, the phenomenon is thought to arise when vibrations allow electrons to form pairs, which stabilizes their path and allows them to flow without resistance. But with so few available electrons in graphene, the fact that they can pair up suggests that the interaction at play in this system is much stronger than what

happens in conventional superconductors.

Graphene-based devices will be easier to study than cuprates, which makes them useful platforms for exploring superconductivity, says Bascones. For example, ‘tuning’ cuprates to explore their different behaviours means growing and studying reams of different samples; with graphene, physicists can achieve the same results by simply tweaking an electric field.

Physicists cannot yet state with certainty that the superconducting mechanism in the two materials is the same. And Laughlin adds that it is not yet clear that all the behaviour seen in cuprates is happening in graphene. “But enough of the behaviours are present in these new experiments to give cause for cautious celebration,” he says. ■

1. Cao, Y. et al. *Nature* <http://dx.doi.org/10.1038/nature26160> (2018).
2. Cao, Y. et al. *Nature* <http://dx.doi.org/10.1038/nature26154> (2018).
3. Ichinokura, S., Sugawara, K., Takayama, A., Takahashi, T. & Hasegawa, S. *ACS Nano* **10**, 2761–2765 (2016).

## POLLUTION

# China tests giant air cleaner to combat urban smog

*Prototype produces clean air and offers an innovative solution to a public-health hazard.*

BY DAVID CYRANOSKI

A 60-metre-high chimney stands in a sea of high-rise buildings in one of China's most polluted cities. But instead of adding to Xian's smog, this chimney is helping to clear the air. The outdoor air-purifying system, powered by the Sun, filters out noxious particles and billows clean air into the skies. Chinese scientists who designed the prototype say that the system could significantly cut pollution in urban areas in China and elsewhere.

The technology has intrigued researchers — especially in China, where air pollution is a daily

challenge. Early results, yet to be published, are promising, says the project's leader, Cao Junji, a chemist at the Chinese Academy of Sciences' Key Laboratory of Aerosol Chemistry and Physics in Xian in central China.

“This is certainly a very interesting idea,” says Donald Wuebbles, an atmospheric scientist at the University of Illinois at Urbana-Champaign, who has heard about the system but not seen it in action. “I am not aware of anyone else doing a project like this one.”

The prototype, built with US\$2 million in funding from the provincial government, has also caught the attention of the president of the Chinese Academy of Sciences, Bai Chunli, who

visited the site last month. Cao says Chinese leaders are eager for solutions to air pollution because it creates such a widespread public-health problem. The Global Burden of Disease Study for 2015, a comprehensive effort to map the world's diseases, found that pollution contributed to 1.1 million premature deaths in China in that year alone.

Cao has submitted a proposal for another tower in Xian, this one 300 metres tall. He is also negotiating proposals with cities in Guangzhou, Hebei and Henan. But the technology has its sceptics, who say that there are much cheaper ways to reduce air pollution.

The concrete chimney sits on top of a



**MORE  
ONLINE**

## IMAGES OF THE MONTH



February's sharpest science shots — selected by *Nature's* photo team [go.nature.com/2fpdhz5](http://go.nature.com/2fpdhz5)

## MORE NEWS

- Colossal family tree reveals environment's influence on lifespan [go.nature.com/2otlaga](http://go.nature.com/2otlaga)
- Ancient genomics studies offer clues to remote Pacific islands' population puzzle [go.nature.com/2fhdjmm](http://go.nature.com/2fhdjmm)

## NATURE PODCAST



Graphene superconductor, and 50 years dreaming of electric sheep [nature.com/nature/podcast](http://nature.com/nature/podcast)

STEVE LOWRY

## FUNDING

# Science wins in Canada budget

*Government focuses its spending on basic research.*

BY BRIAN OWENS

Canadian Prime Minister Justin Trudeau's administration released its 2018 budget on 27 February and scientists couldn't be happier. It includes nearly Can\$4 billion (US\$3.1 billion) in new funding for science over the next five years, a significant portion of which will go to the country's three granting councils. This is in contrast to the Can\$1 billion in new science funding contained in last year's budget — almost none of which went to basic research.

The 2018 budget is “the single largest investment in investigator-led fundamental research in Canadian history”, said finance minister Bill Morneau in remarks to legislators on 27 February.

The Natural Sciences and Engineering Research Council and the Canadian Institutes of Health Research will each receive Can\$354.7 million, and the Social Sciences and Humanities Research Council will get Can\$215.5 million. All three councils will share another Can\$275 million to support research that is “international, interdisciplinary, fast-breaking and higher-risk”.

The move follows recommendations from last year's Fundamental Science Review, a report by an expert panel led by David Naylor, former president of the University of Toronto. He was “relieved and pleased” with this “historic recalibration” in science funding.

Research infrastructure gets Can\$763 million extra over five years, and a pledge of permanent government funding. And early-career scientists receive a further Can\$210 million, also over five years, through a programme that supports researchers at universities across the country.

But scientists didn't get everything they wanted. For instance, there was no new money for the Climate Change and Atmospheric Research programme. Without an influx of cash, several of its research stations in the high Arctic will have to shut down.

Despite that, this budget is a testament to the campaign waged by Canadian researchers over the past year to ensure that the government took the recommendations in the Fundamental Science Review seriously, says Katie Gibbs, executive director of the science campaign group Evidence for Democracy in Ottawa. “It really shows the government spent the last year listening to the community.” ■



Inside a chimney that releases filtered air, part of a pilot project to reduce smog in Xian, China.

DAVID CYRANOSKI/NATURE

large open structure with a glass roof. Solar radiation hitting the glass heats the air, causing it to rise into the tower. The air then passes through a wall of industrial filters before billowing out of the chimney.

“This is a very well-designed and well-made prototype,” says Renaud de Richter, a chemical engineer at the Higher National Institute of Chemistry in Montpellier, France, who has worked on solar-energy towers similar to those that inspired Cao's system. Richter says that Cao's success could help to convince investors to support other applications based on the flow of solar-powered air through chimneys.

Pollution peaks during winter in China, and Cao conducted his first test of the system's air filters over two weeks in January. At the tower, and at 10 monitoring stations across a 10-square-kilometre area, he placed monitors that measured particulate matter less than 2.5 micrometres in diameter (PM<sub>2.5</sub>), a type of pollution that has plagued Chinese cities.

He found that the tower expels between 5 million and 8 million cubic metres of filtered air a day in winter. During the study period, the surrounding air monitors registered a 19% decrease in PM<sub>2.5</sub> concentrations compared with monitors in other parts of the city. Cao is preparing the results for publication.

The project leader says that the prototype's impact was local, so he proposes creating arrays of about half a dozen larger chimneys distributed around urban centres. “We need multiple systems so that significant reduction of air-pollution

concentration can be achieved,” he says.

Neil Donahue, who studies atmospheric particles at Carnegie Mellon University in Pittsburgh, Pennsylvania, says there is little doubt that pulling a large volume of air through high-efficiency particulate filters will clean it. But he wonders if the benefits will be worth the environmental damage caused by building and running such facilities. Turning the same amount of power into clean electricity, or not emitting the pollution in the first place, might achieve the

**“This is certainly a very interesting idea. I am not aware of anyone else doing a project like this one.”**

same pollution cuts, he says.

Wuebbles also worries that the chimney wouldn't filter precursors to particulate matter, such as sulfur dioxide gas and nitrogen oxides, or secondary gaseous pollutants such as ozone. “While the sky may look cleaner, the air quality can still be really awful,” he says.

Cao says that the system already removes nitrogen oxides, one of the major precursors of ultra-fine particles and ozone. He also says that concerns about the economics are overblown. He says the pilot project costs about \$30,000 a year to run. Despite some reservations, researchers including atmospheric scientist Jose-Luis Jimenez, at the University of Colorado Boulder, see an advantage in pursuing the technology. “I'd definitely say it is worth exploring it more, though I am not convinced either way at this point,” Jimenez says. ■





A US weather satellite undergoes testing in a thermal vacuum chamber before launch.

## RESEARCH

# Latest US satellite aids forecasters

*Scientists tackle obstacles to using data in weather models.*

BY JEFF TOLLEFSON

**T**he United States filled a crucial gap in its weather-forecasting arsenal with the launch of its latest geostationary satellite on 1 March. The craft will enable meteorologists to track hurricanes and other threats as they develop. It will also beam down data that researchers can use to measure air temperature and humidity — if they can work out how to incorporate them into their models.

Scientists currently can't use much of the information collected by geostationary satellites, which sit above a particular location on Earth, and polar-orbiting satellites, which swing around the planet's poles. It's a long-standing problem caused by the kind of data collected and the large uncertainties that arise when forecasters try to integrate the measurements into their weather models. Now researchers are starting to overcome these technical challenges, with encouraging results for both short- and longer-term forecasts.

The Geostationary Operational Environmental Satellite-17 (GOES-17) will assume a position above the equatorial Pacific Ocean. When its data are combined with those from the identical GOES-16, which is already parked over the Atlantic Ocean, they will monitor

Earth from Africa to New Zealand. Weather forecasters use such geostationary satellites to track storms, and their models incorporate limited data on atmospheric moisture and wind speed and direction.

"There is this huge treasure trove of information," says Fuqing Zhang, a meteorologist at Pennsylvania State University in University Park. He has experimented with incorporating some of the unused data from satellites into his models, with promising results. "We can show dramatic improvements in weather prediction, but you do need a dedicated research effort," he says.

In a study currently in review at the *Bulletin of the American Meteorological Society*, Zhang shows that integrating high-resolution data from GOES-16 into an experimental weather model bolstered predictions of the early development and intensity of Hurricane Harvey, which struck Texas last August.

Without the extra data, a forecast predicted that the storm would become a category 1 hurricane; in fact, it grew into a category 4 monster before making landfall. Zhang also

included the additional information in a weather model that the US National Weather Service is planning to roll out as early as this year. Those extra data improved forecasts of precipitation amounts and the storm's path.

Incorporating such information into the models has been difficult in part because geostationary data provide fewer measurements for any given vertical slice of the atmosphere than do polar orbiters, which circle Earth at lower altitudes. That means researchers have less information and higher uncertainties when it comes to translating the data into measurements that the models can use, such as air temperature and humidity.

"It's not trivial," says Dan Lindsey, a research meteorologist with the US National Oceanic and Atmospheric Administration in Fort Collins, Colorado, who works on the current GOES satellites. "You can't just take a satellite image and just shove it into the model."

Meteorologists also struggle to incorporate data on cloudy areas recorded by polar-orbiting satellites. This is because clouds have more-complex microphysics than does the open sky, so even small errors in the models can cascade into large uncertainties in the forecast. And that's the fundamental problem, says Alan Geer, an atmospheric scientist with the European Centre for Medium-Range Weather Forecasts (ECMWF) in Reading, UK. "It's those areas with clouds and precipitation that are associated with the most interesting weather."

The ECMWF has been leading the way in this field for more than a decade, and now incorporates much of the data from cloudy regions taken by polar-orbiting satellites; most major government forecasting centres are now following suit (A. J. Geer *et al.* *Q. J. R. Meteorol. Soc.* <http://dx.doi.org/10.1002/qj.3202>; 2018). Zhang cites an unpublished analysis comparing the European model with the latest US National Weather Service model. The US model performed on average as well as, or better than, the European model when using the full suite of atmospheric data from the ECMWF. But when researchers ran the same model with the usual data from the US forecasting programme, it came up short.

The lesson for the United States is that satellites and models aren't enough, Zhang says. "Our nation has put so much money into launching beautiful satellites, but we haven't really put as much effort into how to put the satellite information into the models." ■

## CORRECTION

The News story 'Rescued radar maps reveal Antarctica's past' (*Nature* **552**, 299–300; 2017) erroneously said that Delft University of Technology in the Netherlands helped to run a radar-mapping programme in Antarctica in the late 1960s. In fact, it was the Technical University of Denmark in Lyngby.





# The cruellest seas

*Extreme floods will become more common as sea levels rise.*

BY ALEXANDRA WITZE

**O**n 8 September 2017, Thomas Wahl checked in at London's Gatwick Airport for a nearly-empty flight to Orlando, Florida. A coastal engineer at the University of Central Florida, Wahl knew what was heading for his hometown: category-5 Hurricane Irma, which had already battered much of the Caribbean. He got on the plane anyway. "It was me, the pilot and a few Disney tourists who just didn't care," he says.

Irma's heavy rains and powerful winds killed dozens of people across Florida. For Wahl, who rode out the storm in his family's one-bedroom apartment, the experience was a rare chance to witness first-hand a phenomenon he has long worried about: extreme sea levels — what happens when storm surges, high tides and waves combine.

Extreme sea-level events can send water pouring over coastal barriers, swamping people's homes and drowning crucial infrastructure. They've happened, for example, in New Orleans in Louisiana and the surrounding region — still recovering from more than US\$100 billion in damages caused by Hurricane Katrina in 2005 — and in Jacksonville, Florida,

where Irma swamped parts of the city under 2 metres of water, trapping residents and closing bridges and the city's international airport.

Globally, mean sea level is rising by just over 3 millimetres a year, as glaciers and ice caps melt and warming ocean water expands. Researchers have typically focused on understanding the causes and rate of that rise. But swelling seas are also expected to affect extreme sea levels, with devastating effects. In the coming decades, 100-year floods — those that have a 1% chance of hitting in a given year, or an average return interval of 100 years — could occur as often as every year or two (see 'The coming floods'). Across Europe, the cost of coastal flooding could rise by more than a factor of 20 by the year 2100 (ref. 1). And in some regions, the intensity of what constitutes a 100-year flood will become much more severe.

Wahl and a small band of colleagues say that more scientists need to pay attention to the shifting nature of these calamitous events and how they will affect those living near the coast. Such floods will be one of the biggest

Rare events such as the 'bomb cyclone' that battered the US northeast in January are hitting more often.

SCOTT EISEN/GETTY



threats that humanity faces in the future, he says. “When we talk about flood risk, at some point we have to deal with extreme analysis. It’s those high-impact, low-probability events that we really have to worry about.”

By combing historical records and using models to estimate the risk<sup>2</sup>, Wahl and others are making strides in predicting the dangers of such events. The conclusions vary with location. Some coastal communities will face a dangerous rise in the number of extreme sea-level events. Others are likely to be more prone to ‘nuisance’ flooding — inundations that swamp streets and make life difficult for residents but have less-dramatic overall effects.

Communities will need such knowledge to help them prepare, says Maya Buchanan, a climate-change specialist at the consulting firm ICF International in New York City. Officials can tackle nuisance flooding by improving drainage and other infrastructure. But extreme events call for bigger efforts: building dykes or beefing up sea walls. All told, some 300 million coastal residents are at risk of these events. “This is important for decision-making and society writ large,” Buchanan says.

### COMBING RECORDS

On 6 February 1978, a blizzard of historic power descended on New England. Heavy snow stranded motorists. High tides combined with storm surge to toss coastal homes around as if they were dolls’ houses. All told, 54 people were killed and thousands of buildings destroyed.

The official tide-gauge record from Boston Harbor shows that water levels, not accounting for tides, rose roughly a metre in 12 hours, taking it to one of the highest high-water marks ever recorded there. But it is just one of many records gathered by gauges around the world, which capture the ebb and flow of daily tides, as well as storm-driven surges in water level.

In 2009, a team led by Philip Woodworth of the National Oceanography Centre in Liverpool, UK, decided to assemble as many of these records as possible into a custom-built global database. The team focused on measurements that were made at least once an hour, frequently enough to accurately capture the high-water mark during a rapidly changing storm.

The database, known as the Global Extreme Sea Level Analysis (GESLA) project, has become the go-to source for analysing how extreme sea levels have changed over time. It shows<sup>3</sup> that, since 1970, the magnitude and frequency of extreme sea levels have increased throughout the world. For example, the height of what constitutes a 50-year flood event has risen by more than 10 centimetres per decade in some places<sup>4</sup>.

Most of the blame lies with the rise in mean sea level. As the oceans lap higher and higher against coastlines, incoming storm surges can flood to record high-water marks more easily. One estimate suggests that sea-level rise caused more than \$2 billion of the nearly \$12 billion in damage done to New York City by Hurricane Sandy in 2012.

Other factors also influence extreme sea levels. Long-term atmospheric circulation patterns play a part; a strong El Niño, for instance, pushes water masses around in ways that increase the probability of high sea levels for the US west coast and decrease it for the tropical western Pacific. Changes in the relative height of the land and sea are important, too; much of Scandinavia’s coast has been slowly rising since heavy glaciers disappeared at the end of the last ice age. In southern Asia, the Ganges–Brahmaputra delta is sinking as its sediments compact.

GESLA, which was updated in 2016, now contains 1,355 records from around the world, with more than 39,000 station-years (the number of stations multiplied by the length of their records) of data<sup>5</sup>. Most date from the second half of the twentieth century. But that’s not long enough for researchers who want to improve the long-term statistics. A rule of thumb holds that the frequency of events can be extrapolated up to four times as far into the future as the observational record reaches backward. A

few decades of data won’t be sufficient to inform forecasts for a 10,000-year flood event, as is required for some communities and for sensitive infrastructure such as nuclear power plants.

At Portland State University in Oregon, oceanographer Stefan Talke has been focusing on extending the historical records of water levels in the United States. The US National Oceanic and Atmospheric Administration (NOAA) maintains tide gauges for much of the country. Its records include most of the twentieth century, but can date back to the nineteenth. With colleagues and students, Talke travels to archives around the country, looking for information on tidal patterns and storm surges that NOAA hasn’t systematically digitized. “We have all these questions about what’s going to happen in the future,” Talke says. “If we don’t even understand the past as well as we could, how can we project into the future?”

Talke and his colleagues have ploughed through reams of handwritten data tables and accompanying notes that describe how the measurements were taken. The notes were crucial to assessing the quality of the data; they describe clocks breaking down, ice-clogged gauges and a drunken observer making questionable measurements. All told, the researchers recovered around 300,000 documents, representing more than 6,500 station-years of lost or forgotten measurements<sup>6</sup>.

In Boston, for instance, they found and digitized 50 years of records from before the 1921 start of the modern NOAA record. From that and even older records they calculated that the sea level in Boston has risen, relative to the land level, by 28 centimetres since the 1820s. With that rise, extremes have become routine: what was a 100-year event in the 1820s is now more like an 8-year event.

### BACK IN TIME

The researchers also discovered extreme events that help put more-modern flooding in perspective. A storm they uncovered in 1909, for instance, turned out to have produced as much flooding as the blizzard of 1978. “The archival research shows that the 1978 event really isn’t that anomalous,” Talke says.

But both those high-water marks were exceeded by a ‘bomb cyclone’ in Boston in January, which gained power from rising sea levels to breach sea walls and pour icy waters into neighbourhoods. And last week, another powerful winter storm caused near-record flooding in the Boston area, remarkably just two months after the last record-breaker.

Talke shares his findings with other scientists and with officials at agencies such as the Army Corps of Engineers, which is in charge of federal coastal engineering. He hopes that, by understanding the long-term trends, society will be better equipped to

make decisions about preparing for future high water levels.

Once they have data about past extreme events, researchers have a couple of ways in which they can predict how often extremes will recur. The simplest method is to use something called a Gumbel distribution. This approach was used in the most recent Intergovernmental Panel on Climate Change report on sea-level rise to calculate how often floods would recur under various scenarios of greenhouse-gas emissions. But it is relatively simplistic and does not do a good job of capturing extreme events, Wahl says. For instance, a typical Gumbel analysis might look at the highest annual water level for a given location. That means that only the highest flood counts for a year that might have had more than one big storm.

Buchanan and her colleagues recently adopted a different approach, the generalized Pareto distribution, to incorporate all hourly water-level observations above the 99th percentile. That means more data points, and therefore a more accurate picture of the variation over time. The group studied all NOAA tide-gauge records that went back at least 30 years, then combined those data with an analysis of sea-level rise to predict how often floods would

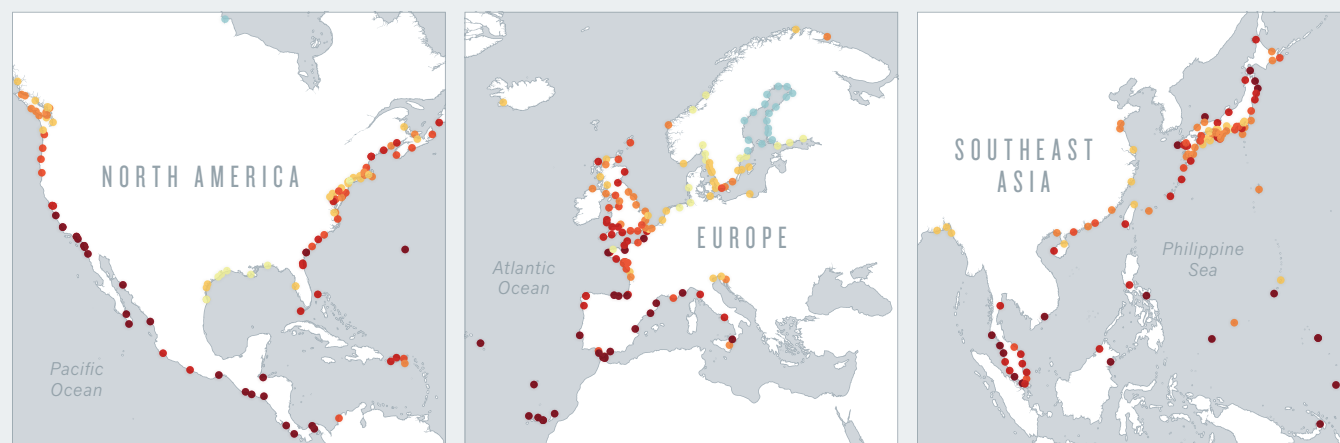
*“At what point are we looking at the 50-year water level being exceeded every single year?”*

## THE COMING FLOODS

Mean sea levels are rising around the globe, which is affecting the rate of floods in various locations. By 2050, some places (darker red dots) can expect to see what is currently considered a 100-year flood event recur as often as every one or five years on average. In other areas, today's 100-year events will not become more common or may even become rarer (light-blue dots), such as along Scandinavian coasts where the land is rising relative to the sea. Low-lying Pacific islands are among the most vulnerable to increased coastal flooding.



Estimated frequency by 2050 of today's 100-year floods (years) ● 1–2 ● 2–5 ● 5–10 ● 10–20 ● 20–50 ● 50–100 ● 100–10,000



occur in different locations, and how large those floods would get<sup>7</sup>.

The bottom line is simple. “Floods are going to become more frequent,” Buchanan says. But the analysis also revealed that the coasts of the contiguous United States will fare differently. In eastern cities such as New York City and Charleston, South Carolina, it is the nuisance flooding that will become more frequent. By contrast, western cities such as Seattle in Washington, and San Diego in California, should expect more-frequent flooding from extreme events. The west generally has steeper coastal slopes, which have tended to protect residents. But rising sea levels are providing oomph to surpass what had once been a protective barrier.

The disparity between regions can be stark. If sea level rises by half a metre in Charleston, for instance, today's 100-year flood could hit 16 times more often. In Seattle, the rate goes up to 335, making it more like a 4-month event.

At the University of Chicago in Illinois, oceanographer Sean Vitousek has also been working to understand flood risk, using a statistical technique called a generalized extreme value distribution. In a paper last year in *Scientific Reports*<sup>8</sup>, he and his colleagues combined models of global waves, tides and storm surges with sea-level projections to calculate how coastal flooding might increase in the coming decades. They found that a rise in global mean sea level of 10–20 centimetres — which is expected no later than 2050 — would more than double the frequency of extreme sea-level events in the tropics. Hardest-hit will be low-lying Pacific island nations, where sea-level rise accounts for a significant percentage of the variability of typical floods. Places such as Kiribati, the Marshall Islands and the Maldives are not only threatened with being permanently drowned, but are also at risk of regular floods that can ruin water supplies and render land unfit for farming.

Vitousek's work is some of the first on extremes to incorporate changes in waves — along with tides and storm surges. He hopes to improve the statistical methods even more, to better estimate how often such events might happen. “At what point are we looking at the 50-year water level being exceeded every single year?” he asks. “We need to understand how much time we have left to engineer our way out of the problem.”

Predictions of extreme flooding are muddled by uncertainty over how fast greenhouse-gas emissions are likely to rise. In the first-ever projection of how extreme sea levels could affect Europe's coasts, researchers last year calculated<sup>1</sup> that the height of a 100-year flood could increase by between 57 and 81 centimetres by 2100. But that's an average across all of Europe. In the North Sea region, extreme sea levels could rise by nearly a metre with a high rate of emissions. The Portuguese coast and the Gulf

of Cadiz might actually see a decrease in extreme sea levels, thanks in part to a weakening in extreme winds driving storm surges and waves.

The team that conducted the analysis, led by Michalis Voudoukas, an oceanographer at the European Joint Research Centre in Ispra, Italy, has now turned its attention to calculating the economic impacts. Damage from river floods will rise from 0.04% to 0.1% of Europe's gross domestic product by 2100, Voudoukas reported last December at a meeting of the American Geophysical Union in New Orleans. But damage from coastal floods, currently at 0.01%, will rise to between 0.29% and 0.86%. “Coastal flooding becomes one of the most important natural hazards in the future,” he said.

Voudoukas and others in this emerging field are keen to bring their discoveries to decision-makers in coastal communities. In Orlando, for instance, Wahl is part of a newly launched effort to bring together engineers, oceanographers, economists, social scientists and other experts. At the new National Center for Integrated Coastal Research, headquartered at the University of Central Florida, they hope to give policymakers the information they need to figure out how high to build defences, such as dykes or sea walls, in the coming decades.

Knowing just how bad the situation will be — how extreme the extreme sea levels might get — will be a major part of that effort, Wahl says. “I think we can give advice already, based on what we know,” he says.

It's crucial to get started now because adaptation can take a while, Wahl says. Take the Thames flood barrier near London, which has helped prevent flooding but took decades, from its conception after a devastating flood in 1953, to become operational, in 1982. Society simply can't wait that long to begin preparing for the impact of extreme sea levels, Wahl says. “We need to make some decisions now.” ■

**Alexandra Witze** is a correspondent for Nature based in Boulder, Colorado.

1. Voudoukas, M. I. et al. *Earth's Future* **5**, 304–323 (2017).
2. Wahl, T. et al. *Nature Commun.* **8**, 16075 (2017).
3. Menéndez, M. & Woodworth, P. L. J. *Geophys. Res.* **115**, C1011 (2010).
4. Intergovernmental Panel on Climate Change. *Climate Change 2013: The Physical Science Basis* (IPCC, 2013).
5. Woodworth, P. L. et al. *Geosci. Data J.* **3**, 50–59 (2016).
6. Talke, S. A. & Jay, D. A. ‘Archival Water-Level Measurements: Recovering Historical Data to Help Design for the Future.’ In *Civil and Environmental Engineering Faculty Publications and Presentations* 412 (Portland State Uni. Lib., 2017).
7. Buchanan, M. K., Oppenheimer, M. & Kopp, R. E. *Environ. Res. Lett.* **12**, 064009 (2017).
8. Vitousek, S. et al. *Sci. Rep.* **7**, 1399 (2017).



# COMMENT

**BIOSECURITY** Exhaustive study of bioweapons under Vladimir Putin **p.162**



**SCI-FI** Why the book behind *Blade Runner* is more prescient than ever **p.163**

**WOMEN** Gender gaps in astronomy and journals with high impact factors **p.165**

**OBITUARY** Donald Lynden-Bell, galactic-structure pioneer, remembered **p.166**

ROYAL SOCIETY



Crystallographer Kathleen Lonsdale (third from left) at the Royal Society in 1957; she was one of the first female fellows.

## How female fellows fared at the Royal Society

Archive study shows that formal inclusion of women does not automatically lead to their full participation, say **Aileen Fyfe** and **Camilla Mørk Røstvik**.

**I**n 1665, the first issue of the world's longest-running scientific journal appeared: *Philosophical Transactions*. It was not until 1787 that astronomer Caroline Herschel became the first woman to publish a paper in it<sup>1</sup>.

From its beginnings, the journal was tightly linked with the gentlemanly culture of the Royal Society in London<sup>2</sup>. By the 1940s, about 4% of all papers submitted to the Royal Society's journals had a female scientist as an author or co-author<sup>3</sup>. Yet

editorial responsibilities were restricted to scientists who were fellows of the society. So women's involvement in editorial and reviewing roles at the society did not begin until 1945, when the first women were elected as fellows: crystallographer Kathleen Lonsdale and biochemist Marjory Stephenson<sup>4</sup> (see 'Women at the Royal Society'). By 1955, numbers had increased to 10 women — compared with 556 men<sup>5,6</sup>.

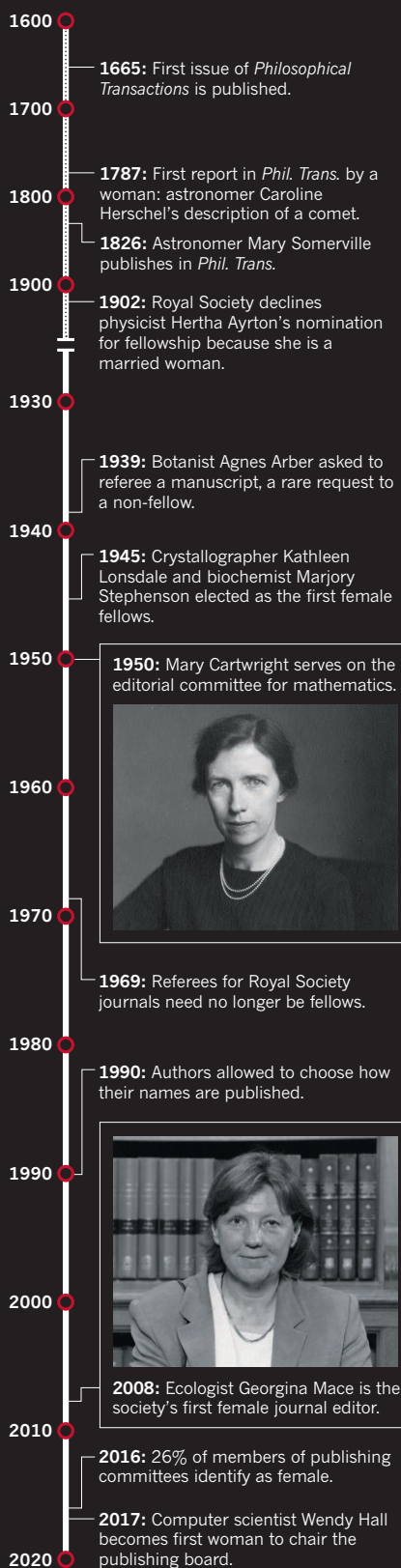
Tempting as it is to point to the 'first woman' as a key moment in institutional

histories, it is surprisingly difficult to see 1945 as marking a significant change in the running of the Royal Society or its publications. This is our conclusion after sifting through decades of archival records, including referee reports, personal correspondence between society officers and referees, and ledgers used to track submitted manuscripts.

Numbers of female fellows of the society did increase over the late twentieth century (see 'Few female fellows'). But the extent ►

## WOMEN AT THE ROYAL SOCIETY

Key moments for female participation in the world's oldest scientific academy.



▶ of women's authorship and editorial work did not follow suit. In fact, in 1955, 2.8% of submitted papers were refereed by a woman; in 1985, only 0.3% were.

### REVEALING REVIEWERS

Modern sociologists who study gender bias in scholarly publication usually focus on the experiences of authors, because the confidentiality of traditional peer review masks the identity of referees and editors<sup>7</sup>. The advent of open peer review enabled a 2017 study<sup>8</sup> to show that women are under-represented as reviewers, and that editors tend to select reviewers of the same gender as themselves.

Our study of the Royal Society's editorial processes adds depth to this discussion. The numbers are tiny, but we know the identities of referees and committee members and can examine their work over several decades.

The first hurdle to having a manuscript accepted in a Royal Society journal was not actually the referee process. Only fellows could officially 'communicate' papers, so would-be authors had to persuade a fellow to act on their behalf. About 80% of papers submitted in the 1950s were from researchers who were not fellows. The gate-keeping system ensured that almost all female scientists submitting to the Royal Society had to do so through a male intermediary.

Of the handful of female fellows who could have acted as communicators, Lonsdale was the most active. Yet she rarely introduced more than one paper a year. And the society archive contains no evidence of her intentionally promoting female-authored manuscripts, although she is known to have invested in the career success of female PhD students<sup>9</sup>.

The fellows who communicated the most papers ran research laboratories, so had a stream of junior scholars working with them. For example, crystallographer Lawrence Bragg, chemist Eric Rideal and physicist Nevill Mott each typically communicated around four or five papers a year in the 1950s. Female lab heads were still relatively rare even several decades later.

Refereeing seems to be a role that could have been more open to female fellows. Again, at the Royal Society, Lonsdale was by far the most active. She wrote 8 of the 10 reports penned by female referees in 1955, and 10 of the 12 in 1956, a level of productivity that made her part of an elite group of active fellows. The majority did little or no refereeing, men and women alike.

Another form of editorial responsibility was sitting on the committees that made final decisions about manuscripts. The first woman to do this was Mary Cartwright, a University of Cambridge mathematician. She joined the mathematics committee in

1950, became the first woman on the society's ruling council (1956–57) and sat on its publications committee (1959–62), which oversaw the committees for individual disciplines. Cartwright was well versed in academic politics, owing to her role as head of Girton College, Cambridge. (Lonsdale also served as vice-president of the society in 1961–62.)

### GENTLEMANLY MICRO-AGGRESSIONS

By the mid-twentieth century, few scientists (or fellows of the Royal Society) were wealthy gentlemen. But gentlemanly codes of conduct still prevailed in academia. Social practices, such as engaging in reasoned discussion at meetings or offering constructive criticism in referees' reports, enabled scholars of different social and intellectual backgrounds to get along (most of the time).

But gentlemanliness was at odds with treating female researchers as peers<sup>10</sup>. University common rooms and faculty clubs were traditionally men-only, as crystallographer Rosalind Franklin discovered on her arrival at King's College London in 1951. (This was not unique to the United Kingdom: women

*"Standard letters and forms used by referees addressed women as 'Dear Sir' until the mid-1960s."*

were not allowed to be full members of the Harvard Faculty Club in Cambridge, Massachusetts, until 1968 — the year before cell biologist Elizabeth Hay

became the first tenured woman at Harvard Medical School.)

In 1923, the Royal Society Club — a private dining group whose members were all fellows — extended a dinner invitation to that year's prize lecturer before realizing that metallurgist and crystallographer C. F. Elam was a woman. She tactfully declined, and accepted "a very beautiful box of chocolates" instead. When crystallographer Dorothy Hodgkin (already a fellow of the society and a Nobel laureate in chemistry) was the prize lecturer in 1972, the club felt "obliged to invite her", according to her contemporary, the physicist Thomas Allibone, even though dinners were normally held in the men-only Athenaeum Club.

'Chivalry' towards female scientists also meant gendered use of titles. In 1960, for instance, an influential fellow wrote a covering note for a paper submitted by crystallographer Helen Scouloudi, describing her as "Miss Scouloudi". He praised her work, but still noted her gender at the expense of her doctorate.

The same thing happened in the society's internal editorial records and published articles. Male authors' names were reduced to initials (unless they were knights). Women who were sole authors had their first names



SOURCE: ROYAL SOCIETY

spelled out, and those who were co-authors were identified by 'Miss' (or 'Mrs') in front of their initials, even if they held PhDs. This means that referees always knew the gender of authors they were reviewing.

This level of care did not extend to women as referees. Standard letters and forms used by referees addressed women such as Lonsdale as 'Dear Sir' until the mid-1960s. At the same time as feminist writer Betty Friedan was describing sexism as "a problem without a name", women's names and titles were being casually neglected at the Royal Society.

In a 2016 interview, a senior member of the publication staff recognized that, in the 1970s, the selection of reviewers depended on 'an old boy network'. He described interactions with female authors (including the naming conventions) as taking a "very gentlemanly approach". He told us, "I remember one female author saying, 'this is discriminatory.' We tended to regard it more as politeness."

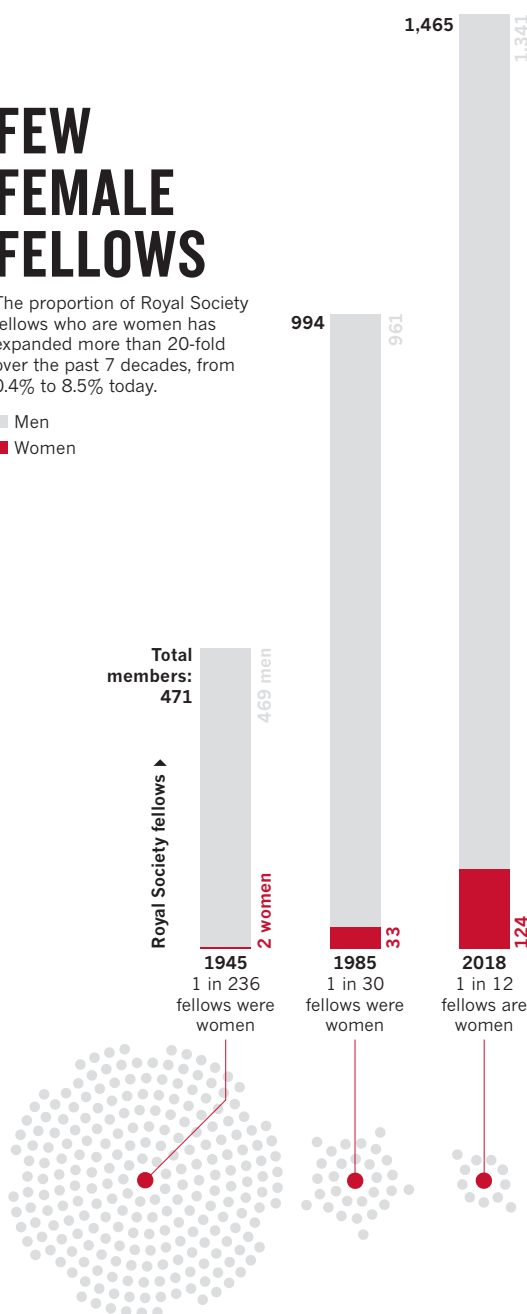
It is hard to say whether referees' awareness of authors' gender affected evaluations: few papers authored by women arrived at the society. In the 1980s, the male referee of an article on chick embryology by two female authors criticized the paper's tone as "too enthusiastic". Another male referee objected that a paper by established palaeontologist Pamela Robinson, which called for evidence of temperature rises and glacial melting to be incorporated into a new approach to palaeoclimatology, was "too ambitious" and used "emotional expressions". (His examples included the phrases "monotonous climate" and "beautiful autumnal colours".) Our sample is limited, but the only instances in which we observed such complaints were about papers authored by women.

By the 1980s, there was a higher proportion of women in the Royal Society than ever before. Yet authors submitting to the society's journals were less likely to have their work considered by women referees. In fact, there are many years for which no woman is listed as a referee at all. This is despite the fact that editorial guidelines were relaxed in 1969 to allow non-fellows to act as referees. Unfortunately, we do not have any information on the number of female scientists who might have been asked to referee, but declined.

## FEW FEMALE FELLOWS

The proportion of Royal Society fellows who are women has expanded more than 20-fold over the past 7 decades, from 0.4% to 8.5% today.

■ Men  
■ Women



Women had taken on other, more visible roles. Biochemist Patricia Clarke and immunologist Brigitte Askonas both served on the council, and developmental biologist Anne McLaren became the society's first elected female officer (as foreign secretary), in 1991. Also, many women elected to the society in the 1980s were already in or near retirement and might have participated less than their younger peers.

Much changed at the Royal Society around 1990, including the abolition of the need for authors to find a 'communicator', and the discriminatory naming conventions. The society's editorial records also moved from physical ledgers to an early computer system, the obsolescence of which has so far prevented us from extending our study into the 1990s.

Two of the society's journals have now had

female editors (ecologist Georgina Mace and geneticist Linda Partridge at *Philosophical Transactions B*, and science historian Anna Marie Roos at *Notes and Records of the Royal Society*). And 26% of members of publishing committees identified as female in 2016 (ref. 11). We do not have data on women's participation in refereeing since 1990.

By February 2018, the number of female fellows of the Royal Society was only 124, or 8.5%. For comparison, 24% of professors in the United Kingdom were female in 2017. Ratios this skewed mean that each elite female scientist faces greater pressure to shoulder more responsibilities — plenaries, panels, mentoring and so on — than her male peers.

A male fellow reflected on this in an interview in 2017. Delighted that a woman was about to chair the society's publishing committee, he told us: "It's generally easier to twist a bloke's arm than to twist a woman's. I'm always reluctant to twist a woman's arm ... It's hard enough being a woman in science most of the time, without taking on all sorts of pro-bono jobs." For us, this illustrates how lingering chivalry may limit women's participation.

We began our research expecting to see a steady rise in women's participation in editorial work after the election of the first women into the Royal Society. Our analysis shows that the admission of women was not, in itself, enough to change the organizational culture of scientific publishing there.

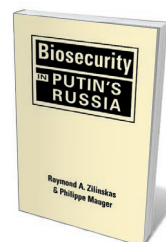
This finding challenges the assumption, often made by powerful institutions, that accepting women into a male-dominated group is enough to bring about equality. Overcoming centuries of tradition is difficult, and long overdue. ■

**Aileen Fyfe** is a professor of history and **Camilla Mørk Røstvik** is a Leverhulme Early Career fellow at the University of St Andrews in Fife, UK.  
e-mail: [akf@st-andrews.ac.uk](mailto:akf@st-andrews.ac.uk)

- Herschel, C. *Phil. Trans. R. Soc. Lond.* **77**, 1–3 (1787).
- Shapin, S. *A Social History of Truth: Civility and Science in Seventeenth-Century England* (Univ. Chicago Press, 1994).
- Røstvik, C. M. & Fyfe, A. *Open Libr. Humanit.* <http://dx.doi.org/10.16995/olh.265> (2018).
- Moxham, N. & Fyfe, A. *Hist. J.* <http://dx.doi.org/10.1017/S0018246X17000334> (2017).
- Mason, J. *Notes Rec. R. Soc. Lond.* **46**, 279–300 (1992).
- Mason, J. *Notes Rec. R. Soc. Lond.* **49**, 125–140 (1995).
- Fox, M. F. *Soc. Stud. Sci.* **35**, 131–150 (2005).
- Helmer, M., Schottdorf, M., Neef, A. & Battaglia, D. *eLife* **6**, e21718 (2017).
- Baldwin, M. *Notes Rec. R. Soc. Lond.* **63**, 81–94 (2009).
- Glick, P. & Fiske, S. T. *Am. Psychol.* **56**, 109–118 (2001).
- Royal Society. *Diversity Data Report 2016* (Royal Society, 2017); available at <http://go.nature.com/2os8xoe>



A serviceman from Russia's chemical, radiological and biological weapons defence unit in 2016.



### Biosecurity in Putin's Russia

RAYMOND A. ZILINSKAS & PHILIPPE MAUGER

Lynne Rienner: 2018.

especially dig into issues such as “genetic weapons” (bioweapons aimed at damaging DNA, potentially of specific individuals or groups) and bio-defence research. Their underlying intention throughout seems to be to examine the likelihood that the Russian government is itself willing to engage in banned activities related to biowarfare

agents. The book thus becomes a technical-scientific detective story.

Zilinskas and Mauger cover a lot of ground, from Russia's current biodefence infrastructure to its diplomatic and propagandistic activities in the context of the 1972 Biological Weapons Convention (BWC). There is plenty of suggestive material. They show that, at least according to internal doctrinal documents, Russia's ostensible rejection of new research on bioweapons is equivocal. At the same time, the Putin regime has ramped up its disinformation campaign aimed at insinuating that the United States is not complying with the BWC, thus providing a possible pretext for its own research into banned areas. This is occurring, the authors remind us, against the backdrop of a largely intact biosecurity infrastructure (encompassing Biopreparat, multiple military facilities and other entities). Meanwhile, the civilian biotechnology sector is floundering, and so might become vulnerable to co-option by the military.

Readers expecting a smoking gun (or festering Petri dish) will be disappointed. The authors do not give any information about specific pathogens or tools under development. What they do present is a meticulous, exhaustively researched and extensively cited investigation. The sources on which Zilinskas and Mauger draw range from arcane technical publications (such as a 2008 military tender for infrastructure improvements) to unofficial propaganda, satellite data, official pronouncements and published interviews; one is with the former head of Russia's military Biological Defense Department, Valentin Yevstigneiev.

Zilinskas and Mauger apply innovative methods to routine data. For instance, they cross-reference lists of institutes with publications by scientists at those institutes; this yields illuminating inferences, such as signs that one body might have shifted from above-board scientific publishing to classified work. They also usefully explain the background to treaties and protocols, and conscientiously distinguish between solid fact and their own opinion throughout.

The book has weak points. Although it is

### BIOSECURITY

# Bioweapons in Russia today

Gary Ackerman praises a meticulously researched tome on biosecurity under Vladimir Putin.

Regimes of all types throughout history have sought to harness science for war. As a result, otherwise beneficial technology can become ‘dual-use’. Biological weapons are among the starker examples: research meant to save lives is used to take them. Now, in the run up to elections in Russia, and with concerns mounting about the nation's role globally, biological-weapons specialists Raymond Zilinskas and Philippe Mauger deliver *Biosecurity in Putin's Russia*.

Bioweapons research in Russia and its environs extends back as far as 1928. It took off in the 1970s, for example through the infamous clandestine Biopreparat network.

There, the Soviets weaponized pathogens including the smallpox and Marburg viruses and the anthrax bacterium *Bacillus anthracis*. Zilinskas and Mauger focus on the years 2012–16, when political tensions between Russia and the West intensified markedly. Concerned by apparent shifts in Russia's pronouncements and actions regarding dual-use activities related to biosecurity, Zilinskas and Mauger write that they wish to “move the discussion over Russian compliance concerns to the public sphere”, where an evaluation based on evidence becomes possible.

They investigate — solely through open sources — the current Russian position. They



well-written and engaging, the detail can become ponderous: more than 100 pages are devoted to military and civilian facilities connected to Russian biodefence. A more judicious use of examples, with the remainder relegated to appendices, would have been preferable to repetitive lists.

The book is also short on synthesis. Like the proverbial blind men describing an elephant, the many chapters answer distinct parts of the central question but fail to tell a coherent story. For example, Zilinskas and Mauger do not explicitly link the Russian establishment's apparent growing willingness to research "weapons based on new physical principles" — which is likely to include biological agents — to its increasingly vehement accusations that the United States is engaging in dubious biological research. Instead, the authors' vague policy prescriptions to the US government seem out of place.

Outright allegations might have undermined the authors' carefully marshalled facts and dispassionate analysis. But this indeterminacy is like watching a prosecutor present a stack of circumstantial evidence, then walk out of the courtroom without delivering a closing argument. The authors' case might be circumstantial, but it is a strong one. A forceful concluding chapter — with appropriate caveats about speculation versus fact — might have done the reader a great service. (My guess — and it is just a guess, because there is no hard evidence — is that Russia is capable of working on any pathogen, with any technique, from CRISPR gene-editing to gain-of-function research.)

Ultimately, these are minor quibbles regarding this trove, which will be new to the world outside Russia. The scholarship and cogent analysis in *Biosecurity in Putin's Russia* are formidable, as rigorous as any assessment of the country's biological-warfare capability by the world's best intelligence agencies. The book is overall a fascinating reflection of the complex web of interests and institutions that have converged to drive Russia's current orientation towards biosecurity. As tensions between the West and Russia grow, questions are bound to arise about Russia's capacities and proclivities for biological weapons. Governments, the non-proliferation community, scientists and institutions involved in international collaborative research should begin looking for their answers here. ■

**Gary A. Ackerman** is an associate professor at the College of Emergency Preparedness, Homeland Security and Cybersecurity at the University at Albany, State University of New York. e-mail: gackerman@albany.edu

## IN RETROSPECT

# Do Androids Dream of Electric Sheep?

Ananyo Bhattacharya toasts Philip K. Dick's prescient science-fiction classic as it turns 50.

When science-fiction writer Peter Watts first opened Philip K. Dick's 1968 *Do Androids Dream of Electric Sheep?*, a word caught his eye. It was "friendlily". How had Dick got that past an editor? As Watts told me: "I knew at that point that Dick had to be some kind of sick genius." Further on in the novel are the boldly redundant "disemelevated" and the sublime "kipple" — a word for 'junk' that encapsulates the stuff's sinister tendency to multiply entropically. Only William Shakespeare coined neologisms as brazenly.

**Do Androids Dream of Electric Sheep?**

PHILIP K. DICK  
Doubleday: 1968.

Yet to debate Dick's strengths as a stylist is to miss the point of *Androids*. For, as with much of his oeuvre (44 novels, 121 short

stories and 14 short-story collections), it is ideas that propel the book into the imaginative stratosphere — and inspired director Ridley Scott to craft the masterly 1982 film adaptation, *Blade Runner*.

Many know of the book solely through the film. But *Blade Runner* is only nominally based on the original. Dick's prescience in ►



Philip K. Dick, pictured in 1982.

► *Androids* lies in his portrayal of a society in which human-like robots have emerged at the same time as advances that make people more pliable and predictable, like machines. The film eschews the intricacies of plot that bring this to the fore in the book.

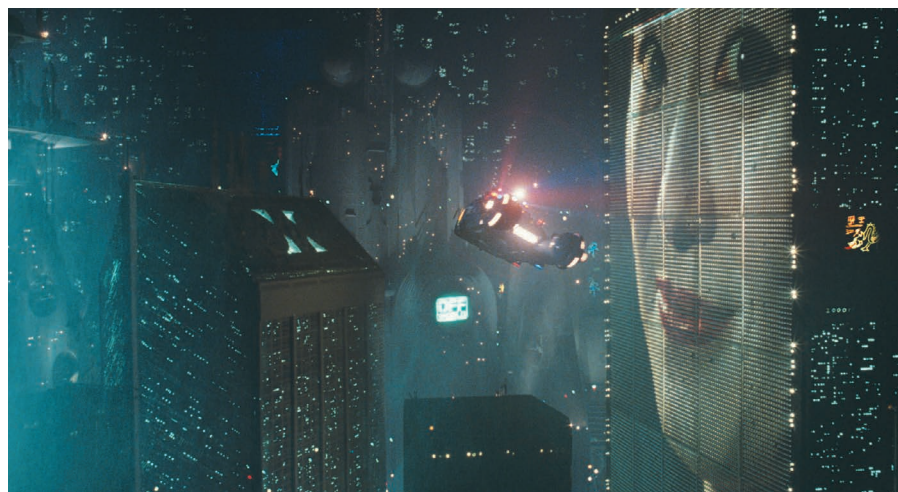
Dick (1928–82) was in many ways as paradoxical as his work. He read widely and was well versed in the science of his day, such as the cybernetics of Norbert Wiener. Yet his formal education ended with school. Shortly after enrolling at the University of California, Berkeley, in 1949 — to study subjects including philosophy — he dropped out, possibly owing to the vertigo and agoraphobia that troubled him throughout his life. The popular image of him, which he encouraged, was of a hallucinogen-addled mystic. But it was amphetamines that fuelled Dick's most heroic bouts of productivity; in 1963–64, he wrote 6 novels in 12 months. His extraordinarily fecund imagination did the rest.

Dick wrote *Androids* in 1966. Others of his books, such as *Ubik* (1969) and his great alternative history *The Man in the High Castle* (1962), were also garlanded with praise. Yet none has perhaps so viscerally affected researchers as *Androids*.

Set in a post-apocalyptic 1992, the book follows bounty hunter Rick Deckard in a risky mission to “retire” (destroy) six state-of-the-art Nexus-6 androids, who have fled to Earth after killing their human masters in a Martian colony. Nexus-6s can be distinguished from humans through the “Voight-Kampff test”. This assesses capacity for empathy, a human faculty that even the most intelligent androids lack.

Deckard embarks on the hunt amid dreams of buying a pet with the reward. Nuclear fallout has extinguished most animal life, and pets are major status symbols. Life-like robotic animals abound, such as the black-faced sheep that Deckard owns; but they are ultimately disappointing. Through caring for an authentic beast, he and his wife Iran hope to transcend the existential fug of living on a planet abandoned by all but the dregs of humanity. Adherents of the religion Mercerism, they feel bound to share such transcendental experiences with others by means of an “empathy box”, a machine that meshes human consciousness.

These days, academic discourse around the work dwells on what distinguishes humans from sophisticated robots — driven by the film. Dick's approach was more nuanced. The name Deckard, for instance, echoes that of seventeenth-century French philosopher René Descartes, who asked whether it was possible to distinguish, without direct access to their minds, a human from an automaton. Deckard explores that ambiguity, wondering uneasily whether he himself is an android. He passes the Voight-Kampff test but, towards the end of the



A still from the 1982 film adaptation *Blade Runner*.

novel, he recognizes a kind of kinship with his quarry. “The electric things have their lives, too,” he says. “Paltry as those lives are.”

Whether such machines should also be accorded rights is a question that researchers wrestle with today. Artificial-intelligence specialist Joanna Bryson, among others, has argued that granting autonomous robots legal personhood would be a mistake: it would render their makers unaccountable. Bryson, an admirer of the book, believes that the mass production of machines with human-like goals and ambitions should be prohibited.

But Dick's chief preoccupation in *Androids* is not the almost-human robot as moral subject. His synthetic beings are inhuman in important ways. They are unable to participate in the rituals of Mercerism, for instance. And their leader, Roy, is a brute who is summarily dispatched. (The film endows him with empathy and even literary flair, saving Deckard's life as he delivers an unforgettable swansong about C-beams that “glitter in the dark near the Tannhäuser Gate”.)

Rather, *Androids* is a meditation on how the fragile, unique human experience might be damaged by technology created to serve us.

The idea that people risk injuring themselves, physically or psychically, by anthropomorphizing machines is not far-fetched. We bond easily with machines. A study last year showed that many people are embarrassed to ask digital assistants such as Apple's Siri questions that betray their own ignorance (S. Kim *et al. Psychol. Sci.* **29**, 171–180; 2017). As far back as the 1990s, electronic pets called Tamagotchis that demanded near-constant care led some owners to neglect important duties. The built-in

compliance of robotic sex dolls currently in development risks eroding relationships.

*Androids* explores this blurred human-machine boundary through Deckard's existential anxiety, and through the “Penfield mood organ”. This device allows humans to dial up urges or emotions, such as “the desire to watch TV, no matter what's on it”, by inputting a number. Named after Wilder Penfield, the real-life twentieth-century neurosurgeon who showed that brain stimulation could elicit sensations and visions, the organ reifies Dick's fear that humans could become more robotic. In this, Dick has been proved spectacularly right. As bioethicist Matt Lamkin has observed, pharmaceuticals that make people happier or more productive — but less contemplative — approximate the mood organ's effect. The smartphone may be the ultimate mood organ: rather than dialling up their own emotions, however, users are increasingly manipulated by the algorithms of tech titans.

To help counter such dehumanizing effects, philosopher Evan Selinger and law scholar Brett Frischmann say that it is time to devise a reverse Turing test. Rather than identifying machines that are indistinguishable from humans, as the original does, the reverse test would determine the extent to which humans remain truly human.

Dick would not have been surprised by any of it. In *Androids*, Iran senses her own blunted emotional response to a life in which caring for machines is the apogee of existence for many, and Earth has been deserted by the smartest. Her answer is to schedule a six-hour bout of self-accusatory depression twice a month. ■

**Ananyo Bhattacharya** is a science correspondent at *The Economist*. His short fiction has been published by *Nature* and in an anthology by *Fantastic Stories*.  
e-mail: ananyobhattacharya@economist.com



# Correspondence

## Academics unite with policy analysts

In our view, academics and funders need extra guidance in working with policymakers (see C. Tyler *Nature* **552**, 7; 2017). This would improve the design of policy-relevant research and help to counter political criticism of academia.

Government ministers rarely have time to build relationships with scientists. Instead, policy analysts in government departments collect information, craft papers for internal discussion and condense these into policy briefs.

These civil servants are typically asked to produce discussion papers on areas outside their expertise within a week (see, for example, M. Howlett and J. Newman in *Policy Work in Canada* 58–76 (Univ. Toronto Press, 2017)). They therefore rely on the most readily available information, including non-peer-reviewed and Internet sources. These might be out of date, flawed or biased. Being able instead to draw on established relationships with scientists would result in policy being developed from a broad and reliable evidence base.

We therefore recommend that academics identify and cultivate relationships with the policy analysts who source raw material for the political machine.

**Marie Claire Brisbois** *Utrecht University, the Netherlands.*

**Kimberly Girling** *Government of Canada, Ottawa, Canada.*

**Scott Findlay** *University of Ottawa, Canada.*

*m.c.g.brisbois@uu.nl*

## Leading journals lack female authors

Our analysis of primary research papers in 15 prestigious multidisciplinary and neuroscience journals in the MEDLINE database indicates that the proportion

of female authors in these journals has been consistently low over the past 13 years. Publication in distinguished journals advances careers, so this under-representation negatively affects the careers of thousands of female scientists.

In *Nature*, for example, women accounted for fewer than 15% of last (senior) authors. By comparison, female scientists received about 27% of prominent research grants from the US National Institutes of Health and from the UK Medical Research Council over the same period.

In these leading journals, we find an impact-factor effect: a negative correlation between the 5-year journal impact factor and the percentage of female first ( $r_s = -0.75$ ,  $P < 0.01$ ) and last ( $r_s = -0.56$ ,  $P < 0.05$ ) authors (for details, see Y. A. Shen *et al.* Preprint at bioRxiv <http://dx.doi.org/10.1101/275362>; 2018).

The proportion of female authors in our set of high-profile journals rose by less than 1% annually in 2005–17. Increasing female representation needs to be a stronger priority (see, for example, *Nature* **541**, 435–436; 2017). Like Microsoft, Google and Walmart, publishing houses have a legal responsibility to avoid discrimination and to implement practices that increase the representation of women and minorities.

**Yiqin Alicia Shen, Yuichi Shoda, Ione Fine** *University of Washington, Seattle, USA.*  
*ionefine@uw.edu*

## Gender gaps in astronomy

Many high-income nations are lagging behind some less-prosperous ones with regard to gender parity in astronomy, according to the International Astronomical Union's (IAU's) latest statistics (see [go.nature.com/2fdji7o](http://go.nature.com/2fdji7o)).

In most wealthy countries, women account for less than

18% of astronomers — including in Switzerland, Germany, the United Kingdom, the United States, Australia and the Nordic nations. Italy (26%), France (25%), Ireland (22%) and Spain (20%) are exceptions. To my knowledge, hardly any women head space agencies such as NASA or the European Space Agency, or lead the editorial boards of astronomy's top journals (see *Nature* **528**, 471–473; 2015).

The proportion of female astronomers is higher in some Latin American and Eastern European countries. Women comprise more than 30% of astronomers in Serbia, Venezuela, Peru, Romania, Bulgaria and Argentina, for example.

Given wealthy countries' reputation for education and outreach, this difference is disappointing. It recalls an age when astronomers were hand-picked royal courtiers and women were excluded. The IAU is taking steps to include more women in its leadership positions (see, for instance, [go.nature.com/2cptoq4](http://go.nature.com/2cptoq4)).  
**Aswin Sekhar** *University of Oslo, Norway.*  
*aswin.sekhar@geo.uio.no*

## Work together for water security

We endorse the proposal for the Indian and Pakistani governments to form a coalition to tackle their transboundary toxic smog (M. Usman *et al.* *Nature* **552**, 334; 2017). A similar approach could be deployed to help resolve the water conflict that has persisted between the two countries since the subcontinent's division in 1947.

The conflict arose because the sources of the rivers flowing into Pakistan are located in India. Despite the signing of a treaty in 1960 that granted Pakistan most of the control over the three western rivers and India full control over the three eastern rivers, tensions have escalated.

These result from inadequate political leadership in the face of increased water demand due to rapid population growth, unpredictable water flow caused by climate change, and dam construction to generate hydroelectricity.

In our view, resources and energy should be directed away from animosity and towards development, with the aim of ending poverty, and improving education and health care.

**Syed Ather Hussain Dow** *University of Health Sciences, Karachi, Pakistan.*

**Vinod C. Nayak** *Manipal Academy of Higher Education, Manipal, India.*

**Ritesh G. Menezes** *Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia.*  
*drsahussain121@gmail.com*

## Electric fish inspire an age of invention

Inspired by the eel's electric organ, Thomas Schroeder and colleagues built a device that provides electricity in a variety of situations (*Nature* **552**, 214–218; 2017, and see [go.nature.com/2hzh4jd](http://go.nature.com/2hzh4jd)). This example of technology derived from a biological concept has echoes of Alessandro Volta's invention of the battery more than two centuries earlier.

Volta (1745–1827) was professor of physics at the University of Pavia in Italy and a fellow of the Royal Society. On 20 March 1800, he sent a letter to Joseph Banks, president of the Royal Society, to communicate his new apparatus (*A. Volta Phil. Trans. R. Soc. Lond.* **90**, 403–431; 1800). He termed this the *Organe électrique artificiel* because it was designed to reconstruct the natural apparatus of electric fish (see also M. Piccolino *Trends Neurosci.* **23**, 147–151; 2000).

**Egidio D'Angelo, Paolo Mazzarello** *University of Pavia, Italy.*  
*dangelo@unipv.it*

# Donald Lynden-Bell

## (1935–2018)

Astrophysicist who predicted that galaxies have black holes at their hearts.

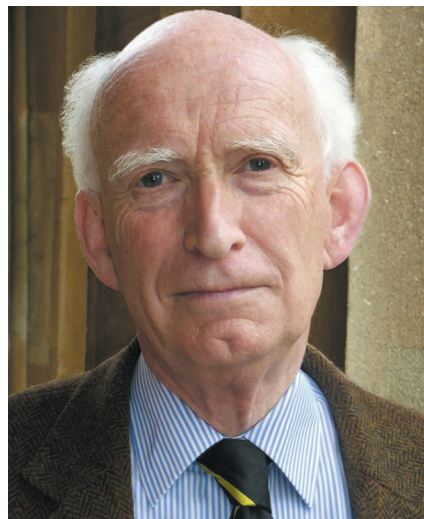
In 1969, Donald Lynden-Bell became the first astrophysicist to suggest that supermassive black holes in the cores of galaxies might generate the profuse energy put out by quasars — the astonishingly luminous distant bodies identified by astronomer Maarten Schmidt earlier that decade. Lynden-Bell proposed that quasars are powered by the release of gravitational energy as material falls into the deep potential well of the black hole, a process that is much more efficient than thermonuclear fusion (D. Lynden-Bell *Nature* **223**, 690–694; 1969).

Over the following decades, he was proved right. We now know that black holes are almost ubiquitous in galaxy cores and seem to have a central role in galaxy evolution. In the past 20 years, the motions of stars at the centre of the Milky Way have revealed a black hole that is four million times as massive as the Sun. And the Hubble Space Telescope has shown that black holes with masses of millions to billions times that of the Sun lie at the core of almost all massive galaxies. Lynden-Bell and Schmidt shared the first Kavli Prize for Astrophysics, in 2008, for their contributions to understanding quasars.

Lynden-Bell died on 6 February 2018. Born in 1935 in Dover, UK, he studied mathematics at the University of Cambridge, followed by a PhD there in theoretical astronomy with Leon Mestel.

In the early 1960s, he spent two formative years at the Carnegie Observatories in Pasadena, California. Using measurements of the composition and orbits of stars taken by Olin Eggen and Allan Sandage, the three developed a model for the formation of the Milky Way, based on the rapid collapse of a large spherical gas cloud (O. J. Eggen *et al.* *Astrophys. J.* **136**, 748; 1962). This was the standard picture for the formation of the Milky Way and other galaxies until the late 1980s, when it was overtaken by the hierarchical-assembly model used today. Lynden-Bell returned to Cambridge in 1962 and moved to the Royal Greenwich Observatory at Herstmonceux, Sussex, in 1965. By this time, he was an astronomer of international stature.

In 1972 he went again to Cambridge, as the first director of the Institute of Astronomy — an amalgamation of the Cambridge Observatories and the Institute of Theoretical Astronomy, which had been founded five years earlier by astronomer Fred Hoyle. The



merger was not initially a happy one, and Donald did not relish his first years at the helm. But he threw himself into new projects, including a plan to build a telescope for the institute (sadly never realized).

He was generous with his ideas and time, and was always curious to know what students were up to, often quizzing them in the corridor. Although he was always supportive, his sharp mathematical insight and booming voice could sometimes be intimidating. He was renowned for taking on young scientists at squash. Student victories were rare.

In the early 1980s, he joined six collaborators in what, at the time, was a huge survey of more than 400 elliptical galaxies. The team — Sandra Faber, her former students Alan Dressler and David Burstein, together with Gary Wegner, Roberto Terlevich, Lynden-Bell and I — formulated a new method for determining the distances to galaxies. Combining this with measurements of how fast the galaxies were moving away, we traced their motions across the sky. It revealed a remarkably coherent flow — with a speed much greater than predicted — in the direction of the constellation Centaurus and close to the plane of the Milky Way, where dust obscures our view of the Universe beyond. Could the corrections used to account for this dust have given rise to a misleading result?

Lynden-Bell was tenacious in scrutinizing these data, and he formulated a test to ensure that the selection of galaxies had not introduced bias. The intense work

generated friction among the team, some of which Lynden-Bell diffused by regaling us with funny stories. On one occasion, he gave a hilarious recitation of the Patrick Barrington rhyme that begins “I had a duck-billed platypus when I was up at Trinity...”

To account for the flow, we hypothesized that there should be many more galaxies behind and beyond the Galactic plane than had been assumed. Dressler nicknamed this concentration the Great Attractor. (Indeed, working with cosmologist Ofer Lahav at around the same time, Lynden-Bell identified a significant over-density of galaxies.) At meetings in 1986, theorists greeted the results with alarm, and observers were sceptical. At a workshop in Santa Cruz, California, astronomer Amos Yahil dubbed our team the ‘seven samurai’ as a nod to our disregard for conventional cosmology.

Lynden-Bell continued to publish influential work on many subjects. These ranged from accretion disks and jets, the violent relaxation of stellar systems, stellar dynamics and spiral structure to general relativity. His extensive studies of the Milky Way and its satellites will be tested in April 2018, when the next tranche of results emerges from the European Space Agency’s Gaia satellite. He wrote several papers on statistical mechanics with his wife, Ruth Truscott, a professor of chemistry at Queen’s University Belfast; they married in 1961 and raised two children, Marion and Edward.

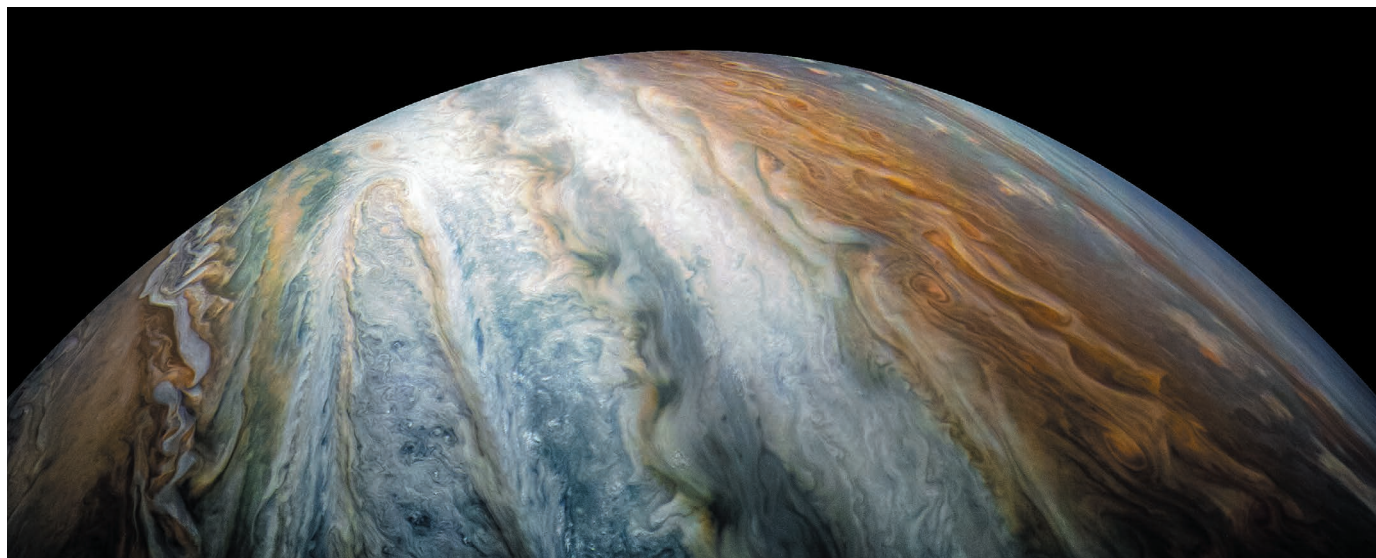
Donald loved sharing the joy and excitement that a life in science had brought him. Fifty years after their first sojourn in California, he and three friends — astronomers Nick Woolf, Wal Sargent and Roger Griffin — returned to the western United States and relived some of the hikes and road trips of their youth. This expedition was made into a 2015 film, *Star Men*, by Alison Rose. Gentle and captivating, it explores comradeship and ageing. Donald travelled around the country to introduce the film and answer questions.

Donald Lynden-Bell was a towering, stimulating, analytical theorist of the sort that is increasingly rare in these days of high-performance computers. ■

**Roger Davies** is professor of astrophysics at the University of Oxford, UK. He worked with Donald Lynden-Bell as part of the seven samurai for a decade from 1979. e-mail: roger.davies@physics.ox.ac.uk

AMANDA SMITH, INST. ASTRONOMY, UNIV. CAMBRIDGE





NASA/JPL-CALTECH/SWRI/MSSS/KEVIN M. GILL

Figure 1 | The surface of Jupiter, as captured by NASA's Juno spacecraft.

## PLANETARY SCIENCE

# A deeper look at Jupiter

NASA's Juno spacecraft has made precise measurements of the gravitational field of Jupiter. The data reveal details of the structure and dynamics of the planet's interior. [SEE LETTERS P.220](#), [P.223](#) & [P.227](#)

JONATHAN FORTNEY

The surface of a planet typically reveals little about the processes at work in the planet's interior. Jupiter's surface consists of alternating bright and dark bands of gas that harbour powerful winds. These winds flow in opposite directions and can reach speeds of more than 100 metres per second. But what happens in the depths below that cannot be seen? In particular, is the planet's interior as dynamic as its surface? In three papers<sup>1–3</sup> in this issue, scientists have used small signatures in the gravitational field of Jupiter to address these questions and to potentially revolutionize our understanding of the internal dynamics of such gas-giant planets.

Jupiter's interior is a dense fluid that comprises a mixture of hydrogen and helium. Energy loss from the interior drives convection currents inside the planet that reach up to the surface. However, neither work in the past few decades on the physics of hydrogen and helium under high pressure, refined measurements of Jupiter's gravitational field from spacecraft nor improved methods to model the planet's structure have been able to determine the mechanics of how the convection operates and whether convective flows in the interior

are related to the banded appearance of the surface (Fig. 1).

One possibility is that the bands are merely a surface phenomenon and that convection in the interior follows an entirely different pattern from convection at the surface. Alternatively, what is seen at the surface could be an extension of deep-seated convective flows that transport energy out of the interior. In both frameworks, sophisticated models have been developed to explain the structure of the bands<sup>4,5</sup>. A main goal of the NASA Juno mission to Jupiter — Earth's nearest gas-giant planet — is to determine which of the frameworks is correct. Because such planets are now known to be common in the Galaxy<sup>6</sup>, achieving this goal would have far-reaching implications for our understanding of this class of astrophysical object.

Iess *et al.*<sup>1</sup> (page 220) tracked the acceleration of the Juno spacecraft in its close elliptical orbit around Jupiter by monitoring the change in frequency, known as the Doppler shift, of radio waves sent back to Earth. Tiny anomalies in these signals revealed details about the mass distribution of Jupiter. Such tracking of Juno was no trivial feat: the authors had to take into account other small accelerations of Juno, including those caused by the absorption and re-radiation of sunlight. They achieved this by

using a sophisticated model of the spacecraft's incoming and outgoing energy.

Iess and colleagues' most stunning finding is that there is a component of Jupiter's gravitational field that does not show north–south symmetry — a peculiar observation for such a fast-rotating gas-giant planet. Kaspi *et al.*<sup>2</sup> (page 223) show that this feature is the result of latitudinal asymmetry in the speed of the winds at the surface. The only way that these winds could affect the planet's gravitational field is if they were relatively deep and involved a substantial amount of mass. This implies that Jupiter's bands are not just a surface phenomenon, thus answering the long-standing question.

Kaspi and co-workers show that the magnitude of the winds decays slowly with depth until about 3,000 kilometres below Jupiter's surface (roughly one-twentieth of the planet's radius), a point at which the pressure is about 100,000 times that of the atmosphere at Earth's surface. The volume of Jupiter in which these winds occur represents about 1% of the planet's mass.

Guillot *et al.*<sup>3</sup> (page 227) confirmed the 3,000-kilometre depth reported by Kaspi and colleagues using the symmetrical component of Jupiter's gravitational field. They demonstrate that, below this depth, the planet's interior rotates as a solid body, despite its fluid nature. This is in accordance with the

prediction that hydrogen ionizes to produce free-moving protons and electrons in such a high-pressure environment. These particles generate strong drag forces that suppress winds flowing in opposite directions<sup>7</sup>.

The three studies confirm previous suggestions that high-precision measurements of a planet's gravitational field can be used to answer questions of deep planetary dynamics<sup>8,9</sup>. In terms of future work, scientists could use the Juno spacecraft to measure the depths of storms on Jupiter such as the Great Red Spot, or to observe the planet's response to tides raised by its large moons. Such analyses would provide a further window into Jupiter's interior.

The work demonstrated here is extremely robust, perhaps unlike other inferences made using data from Juno, including the mass and density of Jupiter's primordial core<sup>10</sup>, that are somewhat model-dependent and rely on our imperfect understanding of the physics of hydrogen under extreme pressure. I do not foresee another leap in knowledge on Jupiter's interior after the Juno mission ends unless astronomers are able to study the planet's internal oscillations<sup>11</sup>, as has been done for the Sun<sup>12</sup>.

Given the inherent complexity of planets, comparative planetary science has become an essential framework through which to study these astrophysical objects. Thankfully, Jupiter has a sibling, the gas-giant planet Saturn. NASA's Cassini mission to Saturn, which ended in 2017, provided a Juno-like data set for Saturn's gravitational field that is now being analysed<sup>13</sup>. Because Saturn has a lower internal pressure than has Jupiter, its atmospheric winds should be able to extend much deeper into its interior before hydrogen ionization and the associated drag forces take control. If a consistent physical picture could be put together for the two gas giants of the Solar System, it would go a long way towards solidifying our understanding of the internal dynamics of this class of astrophysical object. ■

**Jonathan Fortney** is in the *Other Worlds Laboratory, Department of Astronomy and Astrophysics, University of California, Santa Cruz, California 95064, USA.*  
e-mail: jfortney@ucsc.edu

1. Iess, L. *et al. Nature* **555**, 220–222 (2018).
2. Kaspi, Y. *et al. Nature* **555**, 223–226 (2018).
3. Guillot, T. *et al. Nature* **555**, 227–230 (2018).
4. Cho, J. Y.-K. & Polvani, L. M. *Phys. Fluids* **8**, 1531–1552 (1996).
5. Busse, F. H. *Icarus* **29**, 255–260 (1976).
6. Butler, R. P. *et al. Astrophys. J.* **646**, 505–522 (2006).
7. Liu, J., Goldreich, P. M. & Stevenson, D. J. *Icarus* **196**, 653–664 (2008).
8. Hubbard, W. B. *Icarus* **137**, 357–359 (1999).
9. Kaspi, Y. *Geophys. Res. Lett.* **40**, 676–680 (2013).
10. Wahl, S. M. *et al. Geophys. Res. Lett.* **44**, 4649–4659 (2017).
11. Gaulme, P., Schmider, F.-X., Gay, J., Guillot, T. & Jacob, C. *Astron. Astrophys.* **531**, A104 (2011).
12. Christensen-Dalsgaard, J. *Rev. Mod. Phys.* **74**, 1073–1129 (2002).
13. Edgington, S. G. & Spilker, L. J. *Nature Geosci.* **9**, 472–473 (2016).

## HUMAN BEHAVIOUR

# Simple moral code supports cooperation

**The evolution of cooperation is a frequently debated topic. A study assessing scenarios in which people judge each other shows that a simple moral rule suffices to drive the evolution of cooperation. SEE LETTER P.242**

CHARLES EFFERSON & ERNST FEHR

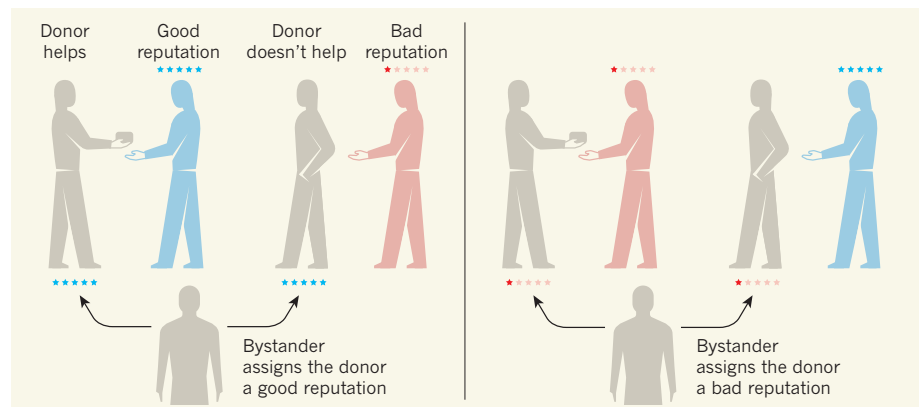
**T**he evolution of cooperation hinges on the benefits of cooperation being shared among those who cooperate<sup>1</sup>. On page 242, Santos *et al.*<sup>2</sup> investigate the evolution of cooperation using computer-based modelling analyses, and they identify a rule for moral judgements that provides an especially powerful system to drive cooperation.

Cooperation can be defined as a behaviour that is costly to the individual providing help, but which provides a greater overall societal benefit. For example, if Angela has a sandwich that is of greater value to Emmanuel than to her, Angela can increase total societal welfare by giving her sandwich to Emmanuel. This requires sacrifice on her part if she likes sandwiches. Reciprocity offers a way for benefactors to avoid helping uncooperative individuals in such situations. If Angela knows Emmanuel is cooperative because she and Emmanuel have interacted before, her reciprocity is direct. If she has heard from others that Emmanuel is a cooperative person, her reciprocity is indirect — a mechanism of particular relevance to human societies<sup>3</sup>.

A strategy is a rule that a donor uses to decide whether or not to cooperate, and the evolution of reciprocal strategies that support cooperation depends crucially on the amount of information that individuals process. Santos and colleagues develop a model to assess the evolution of cooperation through indirect reciprocity. The individuals in their model can consider a relatively large amount of information compared with that used in previous studies.

This increased amount of information is essential for at least two reasons. First, models of direct reciprocity show that having more information allows for many possible strategies, which can paradoxically reduce cooperation<sup>4</sup>. Does something similar happen for indirect reciprocity? Second, indirect reciprocity requires individuals to assess and disseminate reliable information about each other. In a real-world context, this mechanism is most convincing if the amount of information being processed is not excessive. These two considerations suggest that the most compelling models of indirect reciprocity should be simple and should support cooperation in settings in which many alternative possibilities exist.

In Santos and colleagues' set-up, social



**Figure 1 | The stern-judging rule.** Santos *et al.*<sup>2</sup> used a computer-modelling approach to investigate how cooperation might evolve. They investigated scenarios in which a donor can give or refuse help to a recipient depending on the strategy that the donor uses. The donor's action is judged by a bystander who uses a rule (termed a norm) to judge the donor's action and assigns a reputation to the donor that the bystander reports to other members of the society. The authors used this system to test 65,536 different norms in terms of each norm's ability to support the evolution of cooperative strategies. The norm that stood out as being both low complexity and also highly likely to drive the evolution of cooperation is one known as stern judging. This figure shows how the stern-judging norm is used by a bystander to assess a donor's action and thereby assign the donor a good or bad reputation.



interactions involve three individuals: a donor, a recipient and a bystander. The donor uses a strategy to decide whether or not to cooperate and pay a cost that produces a benefit for the recipient. The bystander witnesses this and, using a rule termed a norm, assigns a reputation to the donor that is communicated to others in the population. In future social interactions, this reputation affects whether the donor receives the benefits of cooperation when taking on the role of a recipient.

One version of this interaction is known as a first-order system. In this scenario, two strategies exist. The donor can cooperate or not cooperate (defect). The bystander considers the donor's cooperation or defection when using a norm to assign a good or bad reputation.

Yet even in this simple system, four possible norms exist for the bystander: always assign a good reputation; always assign a bad reputation; assign a good reputation if the donor cooperates and a bad reputation if the donor defects; or assign a bad reputation if the donor cooperates and a good reputation if the donor defects. These norms vary in complexity. The first two are independent of the donor's action and the complexity is low. The latter two norms are dependent on the donor's action and the complexity is relatively high.

This reflects a general pattern. Give a bystander some information, and the level of complexity can vary between the possible norms. Moreover, the complexity of the most-complex norms increases with the information available, and the scope for increasing complexity is striking. In a second-order system, another component is added to the interaction. For example, both the donor and the bystander consider the reputation of the recipient. This allows 4 possible strategies and 16 possible norms. A third-order system could also include the donor's reputation, yielding 16 possible strategies and 256 possible norms<sup>5</sup>.

Santos and colleagues' fourth-order system additionally allows individuals to consider information about the past reputation of either the recipient or the donor. By incorporating the past, a donor's reputation is not dependent on a single point in time. In this scenario, 256 strategies and a staggering 65,536 norms are possible.

With ample scope for complexity in place, Santos and colleagues then examined each norm separately, and allowed the strategies used to evolve (the frequency of use of each strategy could change over time). The strategies that prevail, given a particular norm, affect the amount of cooperation that occurs. One norm, termed stern judging, stands out from the glut of conceivable norms as a relatively low-complexity norm that is highly likely to promote the evolution of cooperation.

The essence of stern judging is to assign a good reputation to a donor who cooperates

with a good recipient or who defects with a bad recipient, and assign a bad reputation to a donor who defects with a good recipient or who cooperates with a bad recipient (Fig. 1). This is a simple second-order norm that supports the evolution of simple and highly cooperative strategies, and it does so even when tested in higher-order systems. From the profusion of feasible norms, more-complex norms do not improve the evolution of cooperation, at least up to the fourth-order system studied by the authors. This suggests that a relatively simple norm, with its correspondingly simple requirements in terms of processing and disseminating information, can suffice to drive indirect reciprocity.

This finding also raises a question for the future. Given so many conceivable norms, why use stern judging? In Santos and colleagues' system, strategies evolve, but norms do not. In reality, strategies and norms evolve together<sup>6</sup>. Both the way people behave (strategies) and the way they evaluate behaviour (norms) change over time, and this process almost certainly involves both genetic and cultural components<sup>7</sup>. Examining the co-evolution of strategies and norms with culture in the mix would be challenging in a fourth-order system, but it would increase our understanding of whether and when we might expect to observe people using reciprocity norms effectively to support cooperation.

In addition, in Santos and colleagues' work, every bystander in a given simulated population uses the same norm. However, in many social settings, there can be variation in the level of subtlety with which different people evaluate social situations. This kind of variation, which could result in bystanders using

norms of different levels of complexity, may or may not<sup>8</sup> result in disagreements between individuals about how to assign reputations. If disagreements occur, how much disagreement can indirect reciprocity tolerate before cooperation breaks down?

Finally, large-scale cooperation occurs in human societies<sup>9</sup>, and efforts to explain how this evolved have generated controversy, possibly because mutually compatible mechanisms are sometimes treated as strict alternatives. Perhaps the next step needed to address this will be to systematically combine multiple mechanisms<sup>4</sup>, including indirect reciprocity, and to test whether specific combinations of mechanisms are especially potent at promoting the evolution of cooperation. ■

**Charles Efferson** is in the Department of Psychology, Royal Holloway, University of London, Egham TW20 0EX, UK.

**Ernst Fehr** is in the Department of Economics, University of Zurich, Zurich 8006, Switzerland. e-mails: charles.efferson@rhul.ac.uk; ernst.fehr@econ.uzh.ch

1. Henrich, J. *J. Econ. Behav. Org.* **53**, 3–35 (2004).
2. Santos, F. P., Santos, F. C. & Pacheco, J. M. *Nature* **555**, 242–245 (2018).
3. Alexander, R. D. *The Biology of Moral Systems* (Routledge, 1987).
4. van Veelen, M., García, J., Rand, D. G. & Nowak, M. A. *Proc. Natl Acad. Sci. USA* **109**, 9929–9934 (2012).
5. Ohtsuki, H. & Iwasa, Y. *J. Theor. Biol.* **239**, 435–444 (2006).
6. Pacheco, J. M., Santos, F. C. & Chalub, F. A. C. C. *PLoS Comput. Biol.* **2**, e178 (2006).
7. Richerson, P. J. & Boyd, R. *Not by Genes Alone: How Culture Transformed Human Evolution* (Univ. Chicago Press, 2005).
8. Uchida, S. & Sigmund, K. *J. Theor. Biol.* **263**, 13–19 (2010).
9. Richerson, P. et al. *Behav. Brain Sci.* **39**, e30 (2016).

#### STRUCTURAL BIOLOGY

## A new era of rationally designed antipsychotics

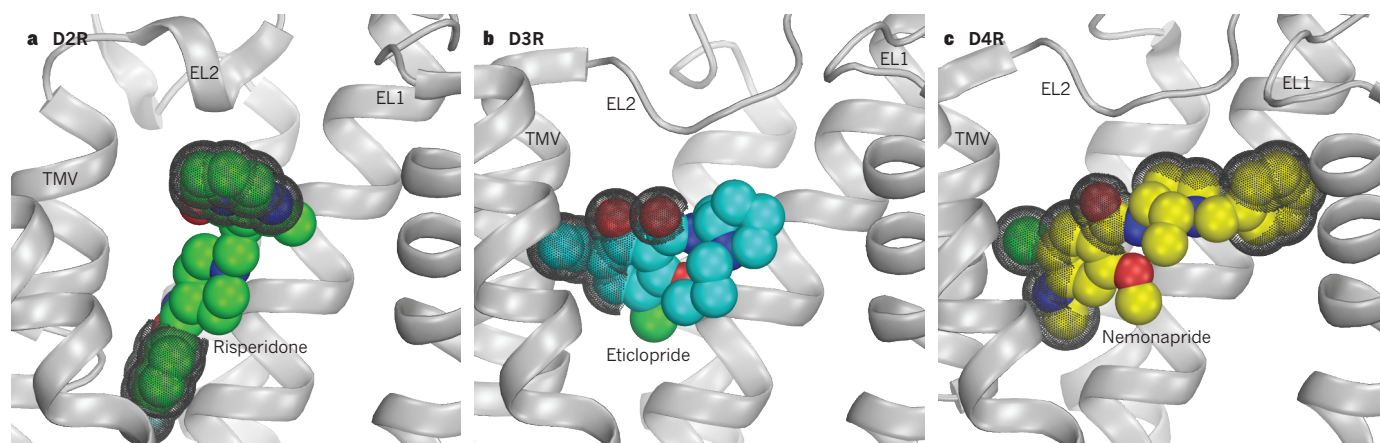
**The ideal drugs for treating schizophrenia are postulated to selectively block the D2 dopamine receptor with optimum binding kinetics. The structure of D2 bound to an antipsychotic sheds light on how to design such drugs. [SEE LETTER P.269](#)**

DAVID R. SIBLEY & LEI SHI

Schizophrenia is a disorder that involves hallucinations, delusions and cognitive impairment, and that affects nearly 1% of the global population<sup>1</sup>. The mainstays of therapy have been drugs that block the activity of the D2 dopamine receptor (D2R), a member of the large G-protein-coupled receptor (GPCR) superfamily of membrane proteins. Unfortunately, most of these antipsychotic

drugs come with a plethora of debilitating side effects, many of which are due to off-target interactions with other GPCRs. On page 269, Wang *et al.*<sup>2</sup> now report the crystal structure of D2R in complex with the antipsychotic drug risperidone. The structure reveals features that might be useful for the design or discovery of drugs that have greater selectivity for D2R than existing therapeutics, and consequently have fewer side effects.

The naturally occurring ligand for D2R is



**Figure 1 | Binding sites within crystal structures of D2-like receptors in complex with drug molecules.** Drugs that block the activity of the D2 dopamine receptor (D2R) are used to treat schizophrenia, but also block the closely related D3 and D4 receptors (D3R and D4R), and exhibit debilitating side effects due, in part, to their interactions with other receptors. **a**, Wang *et al.*<sup>2</sup> report the crystal structure of D2R in complex with the antipsychotic drug risperidone. They observe structural features and drug–receptor binding interactions not observed in the previously reported structure of D3R

with the drug eticlopride<sup>11</sup> (**b**), or of D4R with nemonapride<sup>9</sup> (**c**). The drug molecules are shown as coloured space-filling structures, and the regions enclosed by dots make receptor contacts that are unique to each receptor. The identification of these contacts might help receptor-specific binding pockets to be delineated, which would aid the rational design of receptor-selective drugs. Receptors are shown in grey; thick ribbons are  $\alpha$ -helices; thin regions are unstructured. EL1 and EL2 are extracellular loops. TMV is a transmembrane-spanning segment.

a neurotransmitter called dopamine, which mediates various physiological functions, including the control of coordinated movement, cognition and the reinforcing properties of drugs of abuse. There are five receptors for dopamine, which fall into two subgroups on the basis of their associated intracellular signalling pathways and their affinities for various drugs<sup>3</sup>: D1-like receptors (D1R and D5R) and D2-like receptors (D2R, D3R and D4R). As early as the 1970s, it was hypothesized that the therapeutic effects of antipsychotic drugs were due to them blocking D2-like, rather than D1-like, receptors<sup>4,5</sup>, but the existence of multiple D2-like receptors was not discovered until they were cloned some 15 years later<sup>6</sup>.

Although it has been proposed that antipsychotic-drug action might involve the blocking of D3R or D4R, it is now generally agreed that D2R blockade is necessary, and probably sufficient, for the amelioration of the ‘positive’ symptoms of schizophrenia, such as delusions, hallucinations and disordered thinking<sup>7</sup>. (Antipsychotics currently in use are less effective at treating the ‘negative’ symptoms of this disorder, which include social withdrawal and cognitive impairment.) Progress has been made in the development of D3R-selective<sup>8</sup> and D4R-selective<sup>9</sup> compounds, but there remains a paucity of drugs with high selectivity for the closely related D2R (ref. 10), despite its clear therapeutic importance.

Crystal structures of D3R bound to the drug eticlopride<sup>11</sup> and of D4R bound to the antipsychotic nemonapride<sup>9</sup> have previously been reported. Wang and colleagues’ structure now reveals that risperidone interacts with D2R in a different way from how eticlopride and nemonapride interact with D3R and D4R (Fig. 1). One part of risperidone (known as a benzisoxazole group) extends below the orthosteric site (the

site at which dopamine binds) in D2R, and penetrates deep into a hydrophobic pocket that is not formed in the D3R and D4R structures. A second, extended binding pocket above the orthosteric site in D2R encloses another part of risperidone (a tetrahydropyridopyrimidinone group). This pocket consists of amino-acid residues from extracellular loop 1 (EL1) and three transmembrane helices (TMIII, TMVI and TMVII).

Strikingly, in D2R, a residue within another extracellular loop (EL2), and which is immediately adjacent to an evolutionarily conserved cysteine residue, is buried within the protein and faces the fourth transmembrane helix (TMIV). By contrast, the equivalent residues in D3R and D4R are oriented towards water in the extracellular milieu. EL2 therefore forms a short helical segment in D2R, but is largely extended and unstructured in D3R and D4R (Fig. 1). Consequently, the structural configurations near the EL1 and EL2 interface in D3R and D4R are different from those in D2R.

Wang *et al.* propose that such divergence contributes to the formation of distinct, extended binding pockets in these three receptors, as has been previously suggested<sup>9,11,12</sup>. Drugs designed to selectively engage the distinctive pockets in the D2R structure might display enhanced D2R selectivity. Analogous structure-based drug-discovery efforts have already proved useful in identifying high-affinity compounds<sup>13</sup> that block D3R (ref. 14) or that activate D4R (ref. 9).

Notably, the receptor segments directly above the risperidone-binding site in D2R form a hydrophobic ‘patch’ composed of the side chains of three amino-acid residues, designated Leu94<sup>EL1</sup>, Trp100<sup>EL1</sup> and Ile184<sup>EL2</sup>. This patch potentially restricts the access of molecules to the D2R binding pocket. Wang

and co-workers hypothesized that this feature might regulate the dissociation of risperidone from the D2R binding site, and thus affect its residence time at the receptor.

The authors tested this hypothesis by mutating single residues in the patch and by making a mutant D2R in which both Ile184<sup>EL2</sup> and Leu94<sup>EL1</sup> were replaced. These mutations dramatically reduced risperidone’s residence time from 233 minutes in the wild-type receptor to as little as 6 minutes in the double mutant. This effect is notable because the kinetics of antipsychotic-drug binding to D2R might correlate with a tendency to produce debilitating extrapyramidal side effects (EPS), which include rigidity, tremors and involuntary movements. Antipsychotic drugs that cause fewer EPS, such as risperidone, are said to be atypical, and it has been suggested that antipsychotics with shorter D2R residence times exhibit greater ‘atypicality’<sup>15,16</sup>. Shorter residence times at D2R might enable a minimum level of dopaminergic stimulation, which lessens EPS. The current findings illustrate how elements of the D2R structure can regulate the kinetics of drug binding, which in turn might be associated with desirable therapeutic outcomes.

The hydrophobic patch in D2R is absent in the D3R and D4R structures, presumably because of the separation between the analogous EL1 and EL2 residues in the latter two receptors. Thus, an intriguing question is whether the kinetics of drug binding to D2R are fundamentally different from those to D3R and D4R, particularly for molecules that have similar affinities for the three receptors. In other words, are the kinetics of drug binding to these receptors patch-dependent?

Of further interest is the observation<sup>17</sup> that risperidone is not selective between D2R, D3R and D4R, thus raising the question of how this



drug can bind differently to these receptors and still have identical affinities for them. Additional structures (such as D3R or D4R in complex with risperidone) will probably be needed to answer this. Nonetheless, we expect that Wang and colleagues' D2R–risperidone structure, along with the previous D3R and D4R structures, will accelerate the design and discovery of D2R ligands that have higher selectivity than current antipsychotics, and potentially greater therapeutic impact. ■

**David R. Sibley** is in the Molecular Neuropharmacology Section, National Institute of Neurological Disorders & Stroke,

National Institutes of Health, Bethesda, Maryland 20892-3723, USA. **Lei Shi** is in the Computational Chemistry and Molecular Biophysics Unit, National Institute on Drug Abuse, National Institutes of Health, Baltimore, Maryland 21224, USA. e-mails: sibleyd@ninds.nih.gov; lei.shi2@nih.gov

1. Peralta, J. *et al.* *Arch. Gen. Psychiatry* **64**, 19–28 (2007).
2. Wang, S. *et al.* *Nature* **555**, 269–273 (2018).
3. Keabian, J. W. & Calne, D. B. *Nature* **277**, 93–96 (1979).
4. Creese, L., Burt, D. R. & Snyder, S. H. *Science* **192**, 481–483 (1976).
5. Seeman, P., Lee, T., Chau-Wong, M. & Wong, K. *Nature* **262**, 717–719 (1976).
6. Sibley, D. R. & Monsma, F. J. Jr *Trends Pharmacol. Sci.* **13**, 61–69 (1992).

7. Kapur, S. & Remington, G. *Biol. Psychiatry* **50**, 873–883 (2001).
8. Keck, T. M., Burzynski, C., Shi, L. & Newman, A. H. *Adv. Pharmacol.* **69**, 267–300 (2014).
9. Wang, S. *et al.* *Science* **358**, 381–386 (2017).
10. Moritz, A. E., Free, R. B. & Sibley, D. R. *Cell. Signal.* **41**, 75–81 (2018).
11. Chien, E. Y. *et al.* *Science* **330**, 1091–1095 (2010).
12. Newman, A. H. *et al.* *J. Med. Chem.* **55**, 6689–6699 (2012).
13. Löber, S., Hübner, H., Tschammer, N. & Gmeiner, P. *Trends Pharmacol. Sci.* **32**, 148–157 (2011).
14. Carlsson, J. *et al.* *Nature Chem. Biol.* **7**, 769–778 (2011).
15. Kapur, S. & Seeman, P. *Am. J. Psychiatry* **158**, 360–369 (2001).
16. Sykes, D. A. *et al.* *Nature Commun.* **8**, 763 (2017).
17. Seeman, P. *Clin. Neurosci. Res.* **1**, 53–60 (2001).

This article was published online on 26 February 2018.

## MATERIALS SCIENCE

# Transistors driven by superconductors

**A hybrid transistor device has been made in which a superconductor forms a seamless interface with a semiconductor. The study of such interfaces could open the way to innovative applications in electronics. [SEE ARTICLE P.183](#)**

**YOSHIHARU KROCKENBERGER & YOSHITAKA TANIYASU**

Integrating superconductors with semiconductors has long been thought to be essential to overcome the current limitations of electronic devices, but has been challenging to achieve. On page 183, Yan *et al.*<sup>1</sup> report their use of a technique known as molecular beam epitaxy to grow layers of semiconductors on top of a superconductor. The resulting device has potentially useful electronic properties that hint at future applications for semiconductor–superconductor interfaces.

The development of increasingly sophisticated electronic devices is aided by efforts to make new combinations of materials — or, more specifically, new interfaces between materials, at which potentially useful electronic effects can occur. The credo underlying this concept is that “the interface is the device”<sup>2</sup>. This is particularly true for interfaces involving superconductors.

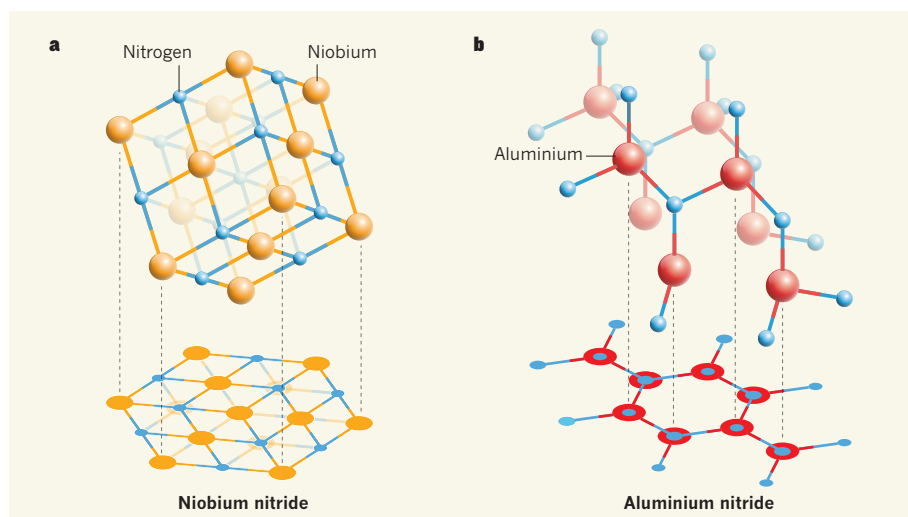
For example, Josephson junctions consist of two superconductors separated by a thin barrier, such as an insulator or a non-superconducting metal. Cooper pairs of electrons — the bound electron pairs that are responsible for superconductivity — can tunnel across the barrier in a fascinating physical process that has led to the development of devices such as those that mix or emit light at terahertz frequencies<sup>3</sup>. Interfacing superconductors with semiconductors<sup>4</sup> such as indium arsenide (an arsenic-based material) can trigger Andreev

reflection processes in which a normal electric current becomes a superconducting current. And if a ferromagnetic material (a material that exhibits the type of magnetism associated with iron) is used as the barrier in a Josephson junction, even more opportunities emerge for the manipulation of controllable electronic states<sup>5</sup>.

Yan and colleagues now report the synthesis of interfaces formed between two

nitrides (nitrogen-containing materials), one a superconductor and the other a semiconductor. Nitride semiconductors<sup>6</sup> are non-toxic, which makes them much more desirable for most applications than toxic arsenic-containing semiconductors. They can be synthesized in well-established procedures using molecular beam epitaxy — a technique in which atomized elements are deposited on a substrate in a vacuum to form thin films of single crystals. Nitride superconductors are also non-toxic, and, more importantly, are highly stable, particularly in ambient conditions (unlike many superconductors). The authors demonstrate that they can fabricate interfaces between a nitride superconductor and devices known as high-electron-mobility transistors<sup>7</sup> (HEMTs) made from nitride semiconductors. HEMTs are widely used in communications infrastructures.

One problem that Yan and colleagues had to contend with is the fact that their nitride semiconductors have hexagonal crystal lattices, whereas the superconductor



**Figure 1 | Aligned views of materials that have different crystal lattices.** **a**, The crystal lattice of the superconductor niobium nitride is cubic, but looks hexagonal when viewed from a particular orientation. **b**, The crystal lattice of the semiconductor aluminium nitride is hexagonal, and can therefore be aligned with the hexagonal arrangement shown in **a**. This allowed Yan *et al.*<sup>1</sup> to prepare electronic devices in which a thin film of aluminium nitride is grown on top of niobium nitride, and the atoms of the two materials are aligned at the interface.

(niobium nitride) is cubic (Fig. 1). This means that the crystallographic symmetry of their devices is broken at the interface between the cubic superconductor and the hexagonal semiconductor. Such broken symmetries can cause unwanted effects at interfaces, and therefore in devices.

This is where the orientation of the superconductor comes into play. Yan *et al.* grew a layer of the cubic superconductor on a substrate so that its lattice was oriented in a way that makes it look hexagonal. To picture this, imagine looking at a dice at an angle in which the diagonally opposite corners are aligned. What you see is a hexagon, even though the dice is cubic.

The same is true of the cubic superconductor on the substrate: a hexagonal arrangement of atoms is exposed on the surface, and the hexagonal semiconductor (aluminium nitride) aligns with this when it forms on top of the superconductor. As a result, the aluminium nitride is not perturbed by broken crystallographic symmetry at the interface, and forms an undistorted layer, as needed for the growth of an HEMT structure. Indeed, the authors observed the formation of certain quantum oscillations in their device; the presence of these oscillations is considered a benchmark of high crystal quality.

Yan *et al.* went on to measure the current–voltage profile of their superconductor–HEMT structure. They observed that this profile of the HEMT is modified by a superconductor-to-metal transition in niobium nitride, and generates a negative differential resistance (NDR) — a property that can be used to increase the power of electrical signals. NDR devices have been known since the end of the nineteenth century<sup>8</sup> and include the Gunn diode<sup>9</sup>, which is widely used to generate microwaves in sensors and measuring instruments. Such devices are of great value for electronic systems that use high-frequency, high-power signals — exactly what is needed in telecommunications networks. In Yan and colleagues' device, NDR can be switched on or off simply by making the temperature lower or higher than the critical temperature for superconductivity (the temperature below which superconductivity occurs).

Combining materials that have different electronic properties without breaking the crystallographic symmetry at the interface is a remarkable feat. However, the mobility of electrons in the device is currently rather low; much higher mobilities can be achieved in devices that use indium arsenide. Achieving mobilities comparable to those of indium arsenide will be extremely challenging. Moreover, the separation between the superconductor and the 2D electron gas — free electrons that are confined to move in only two dimensions — generated in the device will need to be reduced to enable promising quantum effects.

A future goal could be to use the authors' system to generate and observe Majorana

fermions<sup>10</sup> — a type of quasiparticle that would be useful for quantum computing — at the superconductor–semiconductor interface<sup>11</sup>. Charge carriers in electronic devices can be scattered (for example, by crystal defects), and the average time between scattering events needs to be long to stabilize these quasiparticles. Yan *et al.* calculate that the charge-carrier scattering time in their devices is impressively long (66 femtoseconds; 1 fs is  $10^{-15}$  s), but the scattering times will need to be at least 100 times longer, similar to the scattering time in indium arsenide<sup>12</sup>, to stabilize Majorana fermions. It remains to be seen whether this can be achieved in the authors' devices.

Ultimately, Yan and colleagues' work will inspire and accelerate efforts to grow nitride superconductors and nitride semiconductors that enable the ultra-high operating efficiency, structural perfection and opportunities for manipulating electronic properties that have already been achieved in interfaces involving indium arsenide. Because, at the end of the day, the interface is the device. ■

**Yoshiharu Krockenberger and Yoshitaka Taniyasu** are in the Materials Science Laboratory, NTT Basic Research Laboratories, Atsugi, Kanagawa 243-0198, Japan.  
e-mails: yoshiharu.k@lab.ntt.co.jp;  
taniyasu.yoshitaka@lab.ntt.co.jp

1. Yan, R. *et al.* *Nature* **555**, 183–189 (2018).
2. Kroemer, H. *Quasi-Electric Fields and Band Offsets: Teaching Electrons New Tricks* (Nobel Lecture, 8 Dec 2000).
3. Tsujimoto, M. *et al.* *Phys. Rev. Lett.* **108**, 107006 (2012).
4. Kjaergaard, M. *et al.* *Phys. Rev. Appl.* **7**, 034029 (2017).
5. Senapati, K., Blamire, M. G. & Barber, Z. H. *Nature Mater.* **10**, 849–852 (2011).
6. Akasaki, I. *Fascinated Journeys into Blue Light* (Nobel Lecture, 8 Dec 2014).
7. Mimura, T. *IEEE Trans. Microw. Theory Technique* **50**, 780–782 (2002).
8. Frith, J. & Rodgers, C. *Lond. Edinb. Phil. Mag. J. Sci.* **42**, 407–423 (1896).
9. Gunn, J. B. *Solid State Commun.* **1**, 88–91 (1963).
10. Nichele, F. *et al.* *Phys. Rev. Lett.* **119**, 136803 (2017).
11. Krogstrup, P. *et al.* *Nature Mater.* **14**, 400–406 (2015).
12. Shojai, B. *et al.* *Phys. Rev. B* **94**, 245306 (2016).

## EVOLUTION

# Mountains of diversity

**A large-scale analysis of bird diversity and evolution on mountains around the globe explores the relationships between elevation, species richness and the rate of formation of new species. SEE LETTER P.246**

ALEXANDER ZIZKA & ALEXANDRE ANTONELLI

Mountain chains are global centres of biological diversity — they harbour one-third of all terrestrial species<sup>1</sup>. These places have long fascinated biologists<sup>2</sup>, but are notoriously difficult to explore and study. Our knowledge of the distribution of species diversity on mountains is incomplete, as is our understanding of how species richness (the total number of species) and the rates of formation of new species (speciation) vary in single mountain ranges. On page 246, Quintero and Jetz<sup>3</sup> tackle these issues by studying the diversity and evolution of birds on the 46 major mountain ranges of the world.

Mountains can differ substantially in the environment they provide, depending on factors such as bedrock, ruggedness, climatic conditions and the amount of energy available in the region. Moreover, mountains are often far apart, and organisms inhabiting such places can persist in genetically isolated populations owing to factors including terrain complexity and the high variation of habitat types along elevational gradients. Isolated populations often adapt to the local environmental and ecological conditions. When such populations are no longer capable of reproducing with one

another, they form new species<sup>4</sup>. One example of this is the hummingbird *Agelaiocercus kingii*, which is found only in the Andes of South America (Fig. 1).

Quintero and Jetz used large-scale data sets of current distributions of bird species, mined from existing databases and publications, to characterize the relationship between elevation above sea level and species richness. The authors amalgamated data for 9,993 species, representing essentially all the birds that are currently known. Although the patterns observed in different regions vary, the overall trend for most regions is a hump-shaped curve in which species richness is highest at middle elevations, and decreases as elevation increases.

The result confirms findings from previous studies of plants and birds<sup>5,6</sup>. This type of pattern might be driven by the smaller area available for speciation at higher elevations and because the environmental conditions there are more extreme than those on lowlands. For example, large temperature fluctuations between day and night, and an increased exposure to radiation and wind at higher elevations could limit the number of species that can cope with such conditions.

The authors used some innovative approaches for their data analysis. They aimed





**Figure 1 | The hummingbird *Aglaiocercus kingii* in Ecuador.** This species is confined to the Andes mountains of South America. Quintero and Jetz<sup>3</sup> have developed an approach for studying bird distributions on mountains around the world that might help to address how and when biological diversity evolved along elevational gradients.

to capture the 3D structure of biodiversity data by combining elevation and species information. They performed their analyses by grouping the bird data into 'sliced sections' corresponding to trapezoidal prisms that encompass a particular elevation range. This allowed them to assess mountain complexity in a way that improves on conventional ecological methods that often neglect elevation.

Some biodiversity analyses can be affected by issues such as the mid-domain effect, in which a species-richness peak occurs around the centre of a region because of the spatial overlap of species' ranges<sup>7</sup>. The authors developed a subsampling approach that offers a way to address this issue. They counted species, but also factored in the total area that each species occupies when determining species' contributions to species richness. This method also uses a complex randomization procedure that takes a modelling approach to estimate the species' metrics.

Their analysis using this subsampling approach revealed the surprising result that there is a linear decrease in species richness as elevation increases. Nevertheless, one concern is that the subsampled species-richness estimates from this method may not be directly comparable with estimates of species richness calculated in the conventional way — as the total number of species. In addition, the size of each species' range might be a factor linked to its evolutionary history, and therefore relevant for understanding the evolution of mountain species. Additional research might be needed to assess the applicability of this subsampled diversity metric.

Another, perhaps even more interesting finding made by Quintero and Jetz concerns the process underlying the observed species

diversity patterns. The authors used previously estimated<sup>8</sup> information on the evolutionary relationships between the bird species that they studied to calculate the rate of speciation along elevational gradients. They found that this rate is inversely related to subsampled species richness: that is, species are formed at the highest rates where the species richness is lowest, which corresponds to mountaintops. The authors' explanation for this is that environmentally stable lowlands have high diversity, whereas at higher elevations, diversity is governed mainly by frequent immigrations and rapid species replacement during periods of climate change.

A major limitation for studies of biological diversity on mountains is the scarcity of available data. Quintero and Jetz's study uses existing diversity data that have a resolution of at least 110 kilometres horizontally and 500 metres in elevation. This kind of scale can be rather coarse for many mountains, given that environmental and ecological conditions can vary considerably over distances of just a few hundred metres. Although birds are the best geographically documented group of organisms on Earth, with more than 564 million publicly available records (see [www.gbif.org](http://www.gbif.org)), it might come as a surprise that their diversity in many mountains remains poorly documented.

Unfortunately, the geological data of most relevance to biologists are lacking. Quintero and Jetz therefore had to simplify geological complexity in their analyses by using averaged values for key variables, such as the age of mountains. These factors, together with ecological interactions between species, might influence the speciation process<sup>9</sup>, and can vary in a single mountain range.

Speciation rates are also difficult to estimate,

especially over long timescales and for groups, such as birds, that lack a rich fossil record. One potential drawback of the new study is that many relationships between species, and their estimated time of origin, have been calculated on the basis of limited genetic information and with methods that do not take into account the difficulties that sometimes arise during the generation of phylogenetic trees. In some cases, proposed relationships might rely only on comparisons of bird shape and form (morphology) rather than on genetic data.

There is still a long way to go before the phylogeny of birds is fully understood<sup>10</sup>. Large-scale initiatives are under way to sequence the genomes of all bird species as a way to determine more-reliable estimates of the relationships between birds and to improve understanding of their evolutionary history<sup>11</sup>.

Quintero and Jetz's results reveal general and unexpected relationships between elevation, species richness and diversification. Additional data collection in the field by scientists and birdwatchers will be essential and, along with data integration and analysis of the sort spearheaded by Quintero and Jetz, should provide additional insights. It will be particularly interesting to see whether the trends reported by Quintero and Jetz hold true for the rest of the world's species, the diversity and distribution of which are poorly known even at the global level<sup>12</sup> — let alone along elevational gradients on mountains. ■

**Alexander Zizka and Alexandre Antonelli** are at the Gothenburg Global Biodiversity Centre, SE-405 30 Gothenburg, Sweden, and in the Department of Biological and Environmental Sciences, University of Gothenburg. A.A. is also at the Gothenburg Botanical Garden and in the Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts.

e-mail: [alexandre.antonelli@bioenv.gu.se](mailto:alexandre.antonelli@bioenv.gu.se)

- Spehn, E. M., Rudmann-Maurer, K. & Körner, C. *Plant Ecol. Divers.* **4**, 301–302 (2011).
- von Humboldt, A. & Bonpland, A. *Essai sur la Géographie des Plantes* (Schoell, Cotta, 1807).
- Quintero, I. & Jetz, W. *Nature* **555**, 246–250 (2018).
- Hoorn, C., Mosbrugger, V., Mulch, A. & Antonelli, A. *Nature Geosci.* **6**, 154 (2013).
- Kessler, M., Herzog, S. K., Fjeldsø, J. & Bach, K. *Divers. Distrib.* **7**, 61–77 (2001).
- McCain, C. M. *Glob. Ecol. Biogeogr.* **18**, 346–360 (2009).
- Colwell, R. K. & Lees, D. C. *Trends Ecol. Evol.* **15**, 70–76 (2000).
- Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Moores, A. O. *Nature* **491**, 444–448 (2012).
- Condamine, F. L., Antonelli, A., Lagomarsino, L. P., Hoorn, C. & Liow, L. in *Mountains, Climate, and Biodiversity* (eds Hoorn, C., Perrigo, A. & Antonelli, A.) (Wiley, in the press).
- Ricklefs, R. E. & Pagel, M. *Nature* **491**, 336–337 (2012).
- Zhang, G., Jarvis, E. D. & Gilbert, M. T. P. *Science* **346**, 1308–1309 (2014).
- Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. *PLoS Biol.* **9**, e1001127 (2011).

This article was published online on 21 February 2018.

# Meta-analysis and the science of research synthesis

Jessica Gurevitch<sup>1</sup>, Julia Koricheva<sup>2</sup>, Shinichi Nakagawa<sup>3,4</sup> & Gavin Stewart<sup>5</sup>

**Meta-analysis is the quantitative, scientific synthesis of research results. Since the term and modern approaches to research synthesis were first introduced in the 1970s, meta-analysis has had a revolutionary effect in many scientific fields, helping to establish evidence-based practice and to resolve seemingly contradictory research outcomes. At the same time, its implementation has engendered criticism and controversy, in some cases general and others specific to particular disciplines. Here we take the opportunity provided by the recent fortieth anniversary of meta-analysis to reflect on the accomplishments, limitations, recent advances and directions for future developments in the field of research synthesis.**

Synthesizing results across studies to reach an overall understanding of a problem and to identify sources of variation in outcomes is an essential part of the scientific process. Until recently, the results of scientific studies have been summarized in narrative reviews. However, this approach becomes inadequate when there are hundreds of studies on a given research question<sup>1,2</sup>, and the difficulties of carrying out narrative reviews to identify and summarize evidence in a transparent and objective manner have become increasingly apparent as research results have mushroomed across scientific fields<sup>3</sup>.

During the past few decades, scientifically rigorous systematic reviews and meta-analyses, carried out following formal protocols to ensure reproducibility and reduce bias, have become more prevalent in a range of fields<sup>1</sup> (Box 1). Systematic reviews aim to provide a robust overview of the efficacy of an intervention, or of a problem or field of research. They can be combined with quantitative meta-analyses to assess the magnitude of the outcome across relevant primary studies and to analyse the causes of variation among study outcomes (effect sizes). Narrative reviews remain useful for exploring the development of particular ideas (as we do here) and for advancing conceptual frameworks, but they cannot accurately summarize results across studies<sup>4</sup>.

Four decades after its introduction, we are seeing widespread mainstream acceptance of meta-analysis as a research synthesis tool, but also the signs of what may be considered a 'midlife crisis' as it has begun the transition to a mature field. While the number of published meta-analyses has continued to increase rapidly, too many meta-analyses and systematic reviews are of low quality<sup>5–7</sup>. The publication of methodologically flawed meta-analyses indicates that peer reviewers, editors and authors are not fully aware of or are indifferent to the large body of well-developed meta-analytic methodology, and that reviewers might feel unqualified to address statistical issues. Low-quality meta-analyses have attracted strong criticism<sup>5,8</sup> and even calls for a halt in publication of all meta-analyses<sup>9</sup>. Although it is certainly both valid and valuable to criticize poor methodology and reporting, such criticism should result in a call for improved standards (as for pre-clinical trials<sup>10</sup>) rather than abandonment of the field<sup>11</sup>. We believe that the solution lies in the rigorous application of stricter methodological and reporting quality criteria for publishing meta-analyses (see, for example, Tools for Transparency in Ecology and

Evolution; <https://osf.io/g65cb>), and in better training for practitioners and reviewers in the rationales and methodologies of meta-analyses and systematic reviews.

Here we highlight some of the main principles and characteristics of high-quality meta-analytic methodology and briefly summarize the development of the field. We also discuss the limitations, utility and achievements of meta-analysis in several fields and, as a case study, its role in advances in ecology, evolutionary biology and conservation (EEC). Finally, we address several recent criticisms of the meta-analytic approach and suggest ways in which future developments in research synthesis could facilitate the most rapid progress in the fields in which it is used.

## Meta-analyses use well-documented methodologies

Systematic reviews aim to be transparent, reproducible and updatable, and to address well-defined questions. The systematic review process includes the use of formal methodological guidelines for the literature search, study screening (including critical appraisal of eligible studies according to pre-defined criteria), data extraction, coding and often statistical analysis (that is, meta-analysis), along with detailed, transparent documentation of each step. Software, protocols and reporting guidelines for systematic reviews and meta-analyses are well established in many fields; for example, PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses<sup>12</sup>; <http://www.prisma-statement.org/>) is "an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses" and includes a checklist of 27 items and a template flow chart for the presentation of a systematic review (a 'PRISMA flow diagram'; Fig. 1a). Guidelines for developing and preparing systematic review protocols are published in PRISMA-P (<http://www.prisma-statement.org/Extensions/Protocols.aspx>)<sup>13</sup>.

If the systematic review reveals sufficient and appropriate quantitative data from the studies that are being summarized, then a meta-analysis can be conducted. In a meta-analysis, one or more outcomes in the form of effect sizes are extracted from each study. Effect sizes are designed to put the outcomes of the different studies being combined on the same scale, using a suite of metrics<sup>14,15</sup> that includes odds and risk ratios, standardized mean differences, *z*-transformed correlation coefficients and logarithmic ('log') response ratios. It is essential for the effect-size metric used to be

<sup>1</sup>Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794-5245, USA. <sup>2</sup>School of Biological Sciences, Royal Holloway University of London, Egham, Surrey, TW20 0EX, UK. <sup>3</sup>Evolution and Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia. <sup>4</sup>Diabetes and Metabolism Division, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, Sydney, New South Wales 2010, Australia. <sup>5</sup>School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.

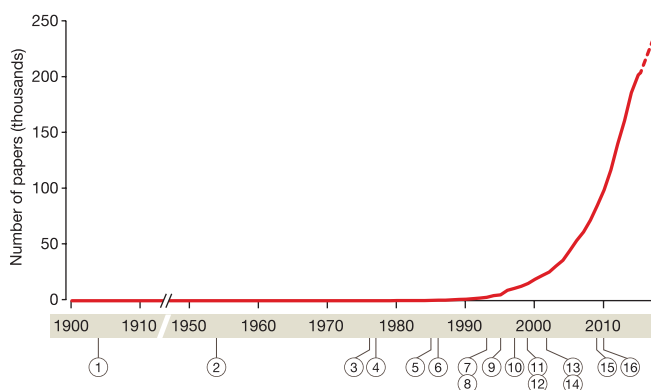


## BOX 1

## A brief history of meta-analysis

The first formal attempt to combine information from multiple sources (see figure) was made in 1904 by K. Pearson<sup>83</sup> with the aim of ascertaining the effectiveness of vaccination in preventing soldiers from contracting typhoid. R. A. Fisher, another important figure in the development of modern statistical science, subsequently introduced a method for combining probabilities from different studies<sup>84</sup>. In the late 1930s, W. Cochran and F. Yates described approaches that were essentially the same as modern fixed-effect and random-effects models<sup>85</sup>, which were later formalized and generalized by Cochran<sup>86</sup>. However, it was not until the insight of psychologists G. Glass and M. Smith in 1977—that outcome measures from different experiments could be standardized and put on the same scale<sup>87</sup>—that meta-analysis began to affect scientific research to a large extent. Meta-analysis was initiated almost simultaneously in medicine and the social sciences<sup>88</sup> and was initially met in all fields with a combination of enthusiasm and condemnation<sup>52,88</sup>. Methodology was formalized and developed in the two decades following 1977 in multiple fields<sup>16,89–91</sup>, with influential studies spreading from medical and social sciences to EEC in the early 1990s<sup>23,92</sup> (Table 1).

Rapid methodological and procedural developments have followed, with cross-disciplinary interactions being the key drivers of progress. The introduction of electronic literature databases and journal articles was central to the development of current practices; a lack of access in poorer institutions and countries hinders scientific progress. The highly interdisciplinary Society for Research Synthesis Methodology (<http://www.srsm.org/>) was established in 2005, after which it began publication of *Research Synthesis Methods*. The large collaborative networks the Cochrane Collaboration (established in 1993; now known as Cochrane; <https://www.cochrane.org>) and the Campbell Collaboration (established in 1999; <https://www.campbellcollaboration.org>) oversee systematic reviews in the medical and social sciences, respectively, bringing practitioners and methodologists together and setting standards for research-synthesis publications and evidence-based guidelines for practice and policy.



- ① 1904 First (medical) meta-analysis published (effect of inoculation against typhoid) (ref. 83)
- ② 1954 First meta-analytic methods formalized (fixed- and random-effects models) (ref. 86)
- ③ 1976 Term 'meta-analysis' coined (ref. 95)
- ④ 1977 First social science meta-analysis published (efficacy of psychotherapy) (ref. 87)
- ⑤ 1985 Statistics textbook dedicated to meta-analytic methods released (ref. 16)
- ⑥ 1986 Method for calculating between-study variance developed (ref. 96)
- ⑦ 1993 Review of 302 social science meta-analyses on treatment efficacy published (ref. 97)
- ⑧ 1993 Cochrane Collaboration established
- ⑨ 1995 Term 'systematic review' introduced (ref. 98)
- ⑩ 1997 Methods for assessing publication bias introduced (funnel plot and Egger's test) (ref. 19)
- ⑪ 1999 QUOROM (Quality of Reporting of Meta-analyses) standards developed (ref. 99)
- ⑫ 1999 Campbell Collaboration established
- ⑬ 2002 Heterogeneity index  $I^2$  proposed (ref. 100)
- ⑭ 2002 Term 'network meta-analysis' coined (ref. 74)
- ⑮ 2009 PRISMA guidelines established (ref. 12)
- ⑯ 2010 *metafor* (free and comprehensive R package for meta-analysis) released (ref. 17)

**Box 1 Figure | Milestones in the history of meta-analysis.** The red line shows the number of papers from a Scopus search; the dashed component indicates the expected future trajectory. The milestone publications<sup>12,16,17,19,74,83,86,87,95–100</sup> are chosen on the basis of two main criteria—precedence and influence (for these criteria, we relied heavily on refs 93 and 94).

readily interpretable, scientifically meaningful and comparable among meta-analyses, and for its sampling distribution to be known, so that statistical models can be constructed appropriately.

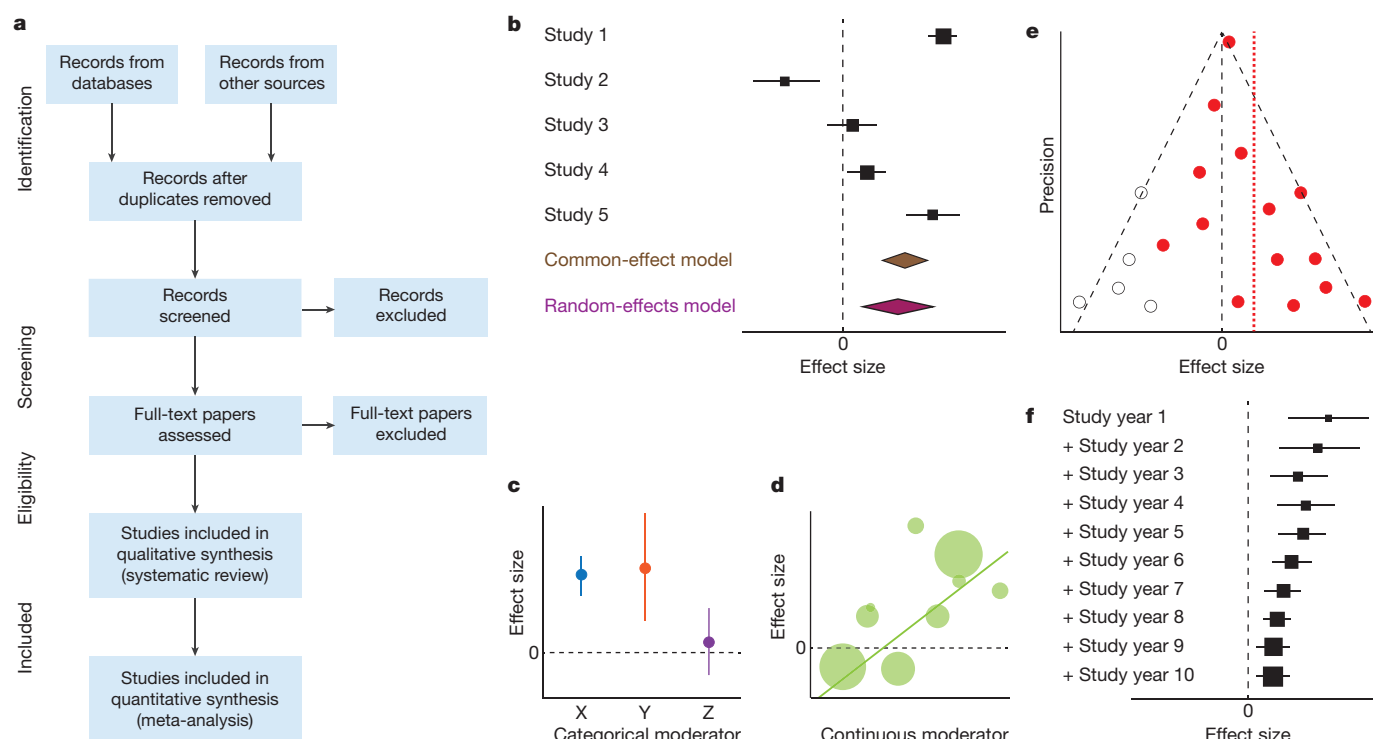
The effect sizes are then entered into a statistical model with the goal of assessing overall effects and heterogeneity in outcomes. These models are based on an assumption of either a common effect ('fixed effect') or random effects (Fig. 1b)<sup>16</sup>. The common-effect (or fixed-effect) model assumes that variation in effect sizes among studies is due to within-study (sampling) variance and that all studies share a common 'true' effect. The random-effects model assumes that, in addition to sampling variance, the true effects from different studies also differ from one another, representing a random sample of a population of outcomes, and is analogous to a random-effects model in an analysis of variance (ANOVA). Thus, random-effects models include an extra variance component to account for between-study variance (heterogeneity) in addition to within-study variance. Common-effect models are based on the assumption that the results apply only to a given group of studies. Random-effects models apply more generally. In carrying out a meta-analysis, the central tendency (the mean) and its confidence limits are evaluated, as well as the heterogeneity in the effect across studies. To identify the magnitude and sources of variation in effect size among studies (Fig. 1c), earlier studies relied on simple heterogeneity tests<sup>16</sup>, whereas more recent work often uses meta-regressions<sup>17</sup>. The 'main effect' or 'grand mean' can be of critical importance or largely irrelevant, depending on the goals of the meta-analysis and the magnitude

and sources of heterogeneity (see sections 'Meta-analysis is essential for progress in science' and 'Meta-analysis in EEC as a case study'). Although these goals differ considerably among disciplines, quantifying heterogeneity is universally important.

Heterogeneity tests and meta-regressions both use weighting based on the precision of the estimate of the effect: larger studies with higher precision are weighted more heavily than smaller and/or more variable studies<sup>18</sup> (Fig. 1b, d). There are many issues to consider in constructing these statistical models, including appropriate weighting and how to account for non-independence (see sections 'Meta-analysis in EEC as a case study' and 'Limitations, controversies and challenges'). In addition, tools have been developed for evaluating publication bias and power and for conducting sensitivity analyses<sup>19–21</sup> (Fig. 1e, f).

### Meta-analysis is essential for progress in science

Meta-analysis has generally been used with two different fundamental goals in mind, resulting in the use of contrasting approaches. The first of these goals is to assess the evidence for the effectiveness of specific interventions for a particular problem or hypothesized causal associations for a condition, often over a relatively small number of studies (fewer than about 25). The second, quite different, fundamental goal is to reach broad generalizations across larger numbers of study outcomes (dozens to hundreds) to provide a more comprehensive picture than can be attained from an individual primary study. The differences in approach



**Figure 1 | Various charts and plots common to meta-analysis.**

**a**, A PRISMA flow diagram<sup>12</sup>, which describes information flow (the number of relevant publications) at the four stages of the systematic review process ('identification', 'screening', 'eligibility' and 'included'). **b**, A 'forest' plot of the various means (symbol centres), confidence limits (95% confidence intervals; whiskers) and precision (indicated by the size or 'weight' of the symbols, with larger symbols indicating greater precision) of the effect-size determined from individual studies (black), and the overall means (various means (symbol centres), and 95% confidence intervals (symbol widths) determined using meta-analysis with a common-effect (or fixed-effect) model (brown) and a random-effects model (purple). This type of plot is used to represent effect sizes and their confidence intervals graphically. **c**, A summary 'forest' plot of the mean effect sizes and 95% confidence intervals for different groups of studies. This type of plot may be used to assess categorical moderators (denoted X, Y and Z here) and

are common in EEC and some social sciences. **d**, A 'bubble' plot showing a line predicted from a meta-regression analysis; the sizes of the bubbles reflect the sample sizes of the individual studies. This type of plot may be used to assess continuous predictors (such as publication year or length of a treatment). **e**, A 'funnel' plot displays the effect size against the precision with which it is estimated, which relates to its weight. Here we illustrate data (red points, with the dotted red line indicating an overall effect) that display 'funnel asymmetry', which could indicate publication bias, along with data (open circles) obtained after applying the trim-and-fill method, a sensitivity analysis that corrects for a potential publication bias. **f**, A 'forest' plot of a cumulative meta-analysis in which outcomes are added into the analysis in chronological order, demonstrating an increase in precision and a convergence of effect sizes as studies are added, and a temporal trend across studies. The dashed black lines in **b–f** indicate 'no effect' of an intervention on the outcome.

and goals affect not only the scale of meta-analyses, but every step of the research synthesis, from study inclusion criteria to the statistical models used. In both approaches, meta-analysis is used to synthesize evidence across studies to detect effects, to estimate their magnitudes and variation and to analyse the factors (covariates or moderators) that influence them.

When the goal is to assess evidence for specific interventions, the focus of meta-analyses is primarily on accurately estimating an overall mean effect, and may include identifying factors that modify that effect. This approach is exemplified by the PICO (population, intervention, comparator, outcome) framework (and its extensions) for formulating questions, in which specification of these elements is central to the purpose of the synthesis<sup>22</sup>, as it is, for example, when assessing clinical effectiveness or the effectiveness of interventions in other disciplines. Question formulation using PICO has been adopted in a wide range of fields, including medicine and the social sciences. Although moderating factors might be important for understanding how the overall effect is influenced by study or population characteristics, meta-analyses for which the primary goal is to estimate the effects of a specific intervention accurately tend to emphasize the consequences of that intervention for a specific population. This type of meta-analysis must clearly and specifically delineate the population in question. Consequently, the results may apply only to that population; for example, the conclusions of a research synthesis of a medical intervention based on studies that included only middle-aged males might not apply to females or to younger males.

In the second case, when the goal is to reach broad generalizations, the population of studies may be large and heterogeneous and, although estimating the main effect of a particular phenomenon or experimental treatment may be important, identifying sources of heterogeneity in outcomes is often central to understanding the overall phenomenon<sup>23</sup>. Meta-analyses undertaken with the aim of reaching broad generalizations deliberately incorporate results from heterogeneous populations so that broad generalizations and the factors that modify them can be examined and tested. This approach is common in the fields of EEC and in some social sciences, in which meta-analyses have been used to address fundamental problems, to weigh the evidence for prominent theories or hypotheses and to consider the generality of common findings, observations or phenomena<sup>23,24</sup>.

Of course, to some extent there is a continuum rather than an absolute dichotomy in meta-analytic approaches, with overlap between disciplines. A limitation of using broad inclusion criteria is the difficulty in adequately accounting for high heterogeneity. A limitation of a reductionist scope and narrow focus is the limited inference that is possible outside of a narrowly specified population or for factors that modify outcomes, whereas the inclusion of a broader definition of the population of interest and potential factors that could affect outcomes might be highly revealing. Both approaches can be limited or even biased. A collection of many narrowly focused reviews of what is essentially the same intervention can generate spurious results, as can the opposite approach of 'fishing' for



significance among many hypothesized explanatory factors or covariates in an excessively broad study.

For both of these basic goals (evaluation of specific interventions or reaching a broad understanding of a general problem), meta-analysis provides a more powerful and less biased means for clarifying, quantifying and disproving (or confirming) assumed wisdom than do conventional approaches<sup>25</sup> including narrative reviews and flawed quantitative methods such as 'vote counts' (see section 'Limitations, controversies and challenges'). Meta-analytic methods have resolved apparently inconclusive data to arrive at a clearer picture, often more rapidly than other approaches. In medicine, meta-analyses can unambiguously assess the effectiveness of particular surgical or pharmaceutical interventions or the statistical significance of hypothesized causal associations. For example, a meta-analysis of 12 clinical studies was able to demonstrate conclusively a clear relationship between maternal obesity and risk of neural tube defects despite considerable variation in the effect sizes reported in individual studies (from a slightly greater incidence of these birth defects for overweight mothers compared to normal-weight mothers, to three times the risk (odds ratio of 3.11) for severely obese mothers compared to normal-weight mothers)<sup>26</sup>. Similarly, primary studies of the value of a family-based intervention approach for serious juvenile offenders called multi-systemic therapy were seemingly inconsistent; however, despite the logical and theoretical basis for multi-systemic therapy, a meta-analysis found no significant differences between it and conventional social services in the success of outcomes<sup>27</sup>. Both of these meta-analyses have had ramifications for evidence-based practice.

The most consequential effect of introducing formal research-synthesis methodology has been a profound change in the way scientists think about the outcomes of scientific research. An individual primary study may now be seen as a contribution towards the accumulation of evidence rather than revealing the conclusive answer to a scientific problem<sup>25,28</sup>. There are certainly cases where a single revelatory study has completely illuminated and resolved a major problem; however, in many cases syntheses can provide a more general and complete picture of the evidence than can any individual study. The results of initial studies are too often not confirmed by those of subsequent studies or by syntheses of a body of research. Additional major contributions of the introduction of meta-analysis have been increased attention to reporting standards in primary studies, including full and transparent reporting of data and the recognition that studies that report no significant effect are as potentially interesting and valuable as those that report low *P* values<sup>29,30</sup>.

### Meta-analysis in EEC as a case study

Meta-analysis was first adopted by ecologists and evolutionary biologists some 25 years ago (Table 1) and has had a considerable impact on this research field in both fundamental and applied areas. Meta-analytic approaches in ecology were introduced at around the same time as it became increasingly urgent to provide accurate quantitative assessments, predictions and practical solutions to pressing environmental issues such as biodiversity losses, the increase in invasive species and biotic responses to climate change. Meta-analysis has provided tools for summarizing evidence for these effects, their impacts and the effectiveness of interventions. The increased use of meta-analyses and systematic reviews in conservation and applied ecology has been facilitated by the promotion of evidence-based approaches in this field<sup>31,32</sup>, especially through organizations such as the Centre for Evidence-Based Conservation (<http://www.cebc.bangor.ac.uk>) and the Collaboration for Environmental Evidence (<https://www.environmentalevidence.org>; Table 1).

Applications of meta-analyses and, more recently, systematic reviews in EEC have highlighted major gaps in research<sup>33</sup>, provided assessments of the effects of major environmental drivers (such as climate change<sup>34</sup>) and of the effectiveness of conservation and management strategies<sup>31</sup>, and enabled evaluations of the evidence for ecological and evolutionary theories<sup>35</sup>. Examples of influential ecological meta-analyses include quantifications of the effects of biodiversity on ecosystem functioning and

**Table 1 | Development of systematic reviews and meta-analyses in EEC**

Year	Milestone
1991	First meta-analysis in ecology published <sup>78</sup>
1995	Seminal paper by Arnqvist and Wooster <sup>79</sup> published in <i>Trends in Ecology and Evolution</i> , introducing meta-analysis to many ecologists
1995	National Center for Ecological Analysis and Synthesis established in USA
1997	MetaWin, the first software for ecological meta-analysis created <sup>46</sup>
1999	Special feature on meta-analysis published in <i>Ecology</i> , including an influential paper on statistical issues in ecological meta-analysis <sup>50</sup> and the introduction of the logarithmic response ratio as a metric for effect size <sup>80</sup>
2001	First general review of meta-analysis in ecology published <sup>81</sup>
2003	Centre for Evidence-Based Conservation established in UK
2007	Collaboration for Environmental Evidence created
2008/2009	Seminal papers on phylogenetic meta-analysis published <sup>43,45</sup> and phyloMeta software for integrating phylogeny into meta-analyses released <sup>82</sup>
2011	<i>Environmental Evidence</i> (the official journal of the Collaboration for Environmental Evidence) established
2013	First handbook of meta-analysis in ecology and evolution published <sup>73</sup>
2014	OpenMEE, software for ecological and evolutionary meta-analysis, released <sup>47</sup>
2016	First international conference of the Collaboration for Environmental Evidence, in Stockholm

services<sup>36,37</sup>, which demonstrated that declines in species richness have negative effects on the functioning of ecosystems. It has been found<sup>38</sup> that ecological restoration can reverse environmental degradation and increase biodiversity and the provisioning of ecosystem services in a wide range of ecosystems globally, although not to full recovery compared to reference ecosystems.

Similarly, meta-analytic techniques have provided evolutionary biologists the tools to test key hypotheses based on theories of natural selection, sexual selection and animal social behaviour at unprecedented scales<sup>35</sup>. Examples of prominent evolutionary meta-analyses include assessments of correlations between measures of genetic diversity, fitness and population size<sup>39</sup>. One conclusion is that a reduction in population size due to habitat fragmentation reduces genetic variation, which in turn has a negative impact on fitness in the affected populations.

In EEC, meta-analytic techniques have greatly expanded the ability to construct large-scale overviews of study outcomes—over larger spatial scales, different time periods, multiple systems and a diversity of organisms that are beyond the scope of any one researcher or research group. For example, a global meta-analysis<sup>40</sup> of almost 600 latitudinal gradients in species diversity verified the high degree of generality of the decline in diversity with latitude, but also identified important factors that modify this pattern. Meta-analysis has also been a valuable tool for practitioners in EEC involved in collaborative research who wish to combine original results from experiments carried out across multiple study sites<sup>41,42</sup>.

Unlike clinical medicine and the social sciences, fields in which research focuses on a single species, the multi-species nature of much of EEC research and therefore of meta-analyses has led practitioners to integrate phylogenetic comparative methods with meta-analytic models to take into account potential non-independence among lineages due to shared evolutionary history<sup>43–45</sup>. Non-independence among outcomes due to the variation among sources may be more obvious in EEC than in other fields because of the large size and complex data structure of many meta-analyses in EEC. However, non-independence is a ubiquitous problem for research synthesis in most research fields, and much work remains to be done to better model and account for sources of non-independence.

The structural characteristics of data in EEC and the goals of generality typically result in high heterogeneity. Rather than seeking to explain

all of the heterogeneity among studies, the goal is often to identify key factors of commonality—to detect the signals amid the noise when gaining information about these hypothesized key factors is more important than achieving a clean accounting of all sources of variability. This is a different perspective from that of meta-analyses that focus narrowly on, for example, detecting the efficacy of a specific intervention.

Advances in meta-analysis in EEC have been stimulated by many factors, including learning from practitioners in other disciplines, effective and widespread short courses for students and practising scientists, and the development of software that is tailored specifically to this field<sup>46,47</sup>. Methodological innovations in meta-analytic techniques that have been incorporated or developed in EEC, in addition to phylogenetic approaches, include the meta-analysis of factorial experiments<sup>48</sup>, the introduction and wide acceptance of randomization (permutation) tests in meta-analysis<sup>49</sup>, the early embrace of random-effects and mixed-effects models when they were still highly controversial in other disciplines<sup>50</sup>, and methods for the inclusion of qualitative information such as expert opinions<sup>51</sup>.

The introduction and incorporation of meta-analysis in ecological research have raised similar objections to those raised in other disciplines, and these criticisms and others have been similarly refuted across disciplines<sup>11</sup>. For instance, critics have claimed that the potential for publication bias in the literature (that is, the under-reporting of non-significant results or disconfirming evidence<sup>21</sup>) invalidates the use of meta-analysis. This objection has been refuted by research synthesists in many fields, who point out that when publication bias exists, it presents problems that are not unique to meta-analyses, but affect any attempt to summarize the results of the literature or to reach valid conclusions from it. In another instance, as in the early criticisms of meta-analysis in social sciences<sup>52</sup>, some ecologists have claimed that ecological studies are too heterogeneous to be combined statistically in a meaningful way<sup>9</sup> and that ecology is best served by accumulating a catalogue of case studies<sup>53</sup>. Analogously, the basis for the early objections to introducing statistics to ecology in the mid-twentieth century was the inability to fully account for the uniqueness of individual organisms and the micro-site environmental variation using means and statistical tests. Despite the criticism, the introduction of meta-analysis in EEC has been embraced enthusiastically by the majority of scientists in these disciplines as a ‘remote sensing tool’ that helps scientists to generalize the findings of individual studies to reach a broader understanding<sup>11</sup>, and the number of meta-analyses published in EEC has increased exponentially over time<sup>54</sup>.

### Limitations, controversies and challenges

Despite its current utility and future potential, meta-analysis has various limitations as a tool for research synthesis and for informing decisions. Meta-analyses and systematic reviews can highlight areas in which evidence is deficient, but they cannot overcome these deficiencies—they are statistical and scientific techniques, not magical ones. For example, in a systematic review of the literature on hypotheses for explaining biological invasions, a major gap was found<sup>33</sup> in published studies on invasive species in the tropics, highlighting not only what is known but also what is unknown globally about this problem. Although the existence of such knowledge gaps limits the generality of conclusions that can be drawn from the existing literature, the ability of systematic reviews and meta-analyses to identify these gaps is a strength of these approaches because it directs future primary studies to the areas for which evidence is most needed. Other challenges for meta-analyses and systematic reviews include publication bias and research bias<sup>50</sup>, the latter describing the over- or under-representation of populations, species or systems in the literature, which results in a biased view of the totality. The presence of these issues can be strongly suspected by scientists, but although their magnitude can sometimes be estimated in a meta-analysis<sup>19,20</sup>, it cannot be truly corrected in research syntheses<sup>55,56</sup>. Similarly, a synthesis may be constrained by either selective or incomplete data reporting in primary publications<sup>30</sup>.

One undesirable consequence of the growing recognition and high impact of meta-analysis is an increase in less-than-rigorous applications

of these methods and in the application of arbitrary and less-well-justified methodologies that are sometimes inaccurately referred to as meta-analyses. The use of statistically flawed approaches can lead to erroneous and misleading results that masquerade as serious research syntheses. The term meta-analysis should be applied only to studies that use well-established statistical procedures, such as appropriate effect-size calculation, weighting and heterogeneity analysis<sup>57</sup>, and statistical models that take into account the distinct hierarchical structure of meta-analytic data, or to studies that develop rigorously justified methodological advances of these methods. Unfortunately, the term is often misapplied to any study that uses data from several primary publications, regardless of the rigour of the methodology. Statistically flawed procedures such as vote-counting, which provide only limited information about study outcomes, can be very misleading and have long been discredited, are still used in published papers<sup>6,50</sup>. Vote-counting is a deceptively plausible and appealingly convenient procedure whereby the generality of findings in a group of studies is assessed by counting up the number of significant and non-significant results in individual studies (or by elaborations on this approach). Although it is vulnerable to erroneous inferences and provides unreliable information on the magnitudes or heterogeneity of effects, it persists, zombie-like, returning by the efforts of the naive or determinedly ignorant to haunt the scientific literature. Vote-counting is not a meta-analytic technique, and is not an acceptable basis for meaningfully summarizing research results in published papers.

Meta-analyses that are not weighted by inverse variances are common and often poorly justified, and present different problems. Unlike vote-counts, unweighted meta-analyses can be unbiased and may provide information on the magnitude of the effects<sup>8</sup>. However, in an unweighted analysis, within- and between-study variation cannot be readily separated, and so common- and random-effects models cannot be used and heterogeneity may be difficult to assess properly. Unweighted meta-analysis also increases the influence of small studies<sup>29</sup>, which have often been found to report larger and more variable effects than those reported for larger studies (as a result of the smaller studies being more likely to suffer from random noise, and possibly publication bias). An alternative when variances are unavailable from primary studies is weighting by sample size or other metric, but this method does not incorporate the information that an inverse-variance-weighted analysis provides and can introduce unknown biases. These problems are particularly acute with small sample sizes. One argument that is often made in support of unweighted meta-analysis is that the variances needed for a weighted meta-analysis are frequently unavailable owing to poor data reporting in the primary studies, and it is undesirable to leave studies with missing data out of the meta-analysis. One possible solution is to use one of the various methods that have been developed for imputing or otherwise modelling missing data. And, although data reporting practices are being improved slowly, it may be that many older studies are simply inadequate for accurate quantitative reviews. Another argument for unweighted meta-analysis is that the meta-analysis simplifies to an essentially unweighted analysis when between-study variation is much larger than within-study variation<sup>58</sup>. However, a weighted meta-analysis is required to assess the two types of variation in the first place, and we submit that it would be preferable to report the weighted and unweighted results in such cases.

Another unfortunate outcome of the high impact and growing prestige of meta-analysis<sup>59</sup>, coupled with the use of metrics such as citation numbers and *h*-indices in evaluations of research accomplishments, is an unease among some primary researchers about the fairness and rewards of the scientific process<sup>8,60</sup>. Some have decried reviews as “the black-market of scientific currency”<sup>61</sup>, with calls to replace citations to reviews and meta-analyses with citations of primary studies<sup>61</sup>. Worse, research synthesists in medicine have recently been described as “research parasites”<sup>62</sup> of primary studies and the researchers who conduct them. On the other hand, it could be argued that primary studies without context, comparison or summary are ultimately of limited value. Moreover, methods for research synthesis are not the exclusive province of any one group, but



can be used by primary researchers in their own areas of expertise. The introduction of more explicit guidelines and standards for conducting and reporting meta-analyses could address some of these grievances, and we agree that better methods for citing primary studies in meta-analyses should be implemented to give full credit for the original studies. 'Research parasites' can also serve to increase scientific diversity by adding another 'trophic level', thus improving the functioning of the scientific 'ecosystem'.

### Advances, developments and future promise

Meta-analysis is the grandmother of the 'big data' and 'open science' movements. For hundreds of years, scientists have collected data in individual studies, based on observations and experimentation<sup>63</sup>. The introduction and implementation of meta-analytic techniques was the first large-scale, coordinated effort to collect and synthesize pre-existing data to determine patterns, make predictions, reach generalizations and make evidence-based decisions. Discoveries that have resulted from the analysis of big data, in parallel with the development of open-science practices, transparency and the importance of replication of research, are transforming many research areas. 'Big data' refers to large, complex datasets that may be mined for patterns or for making predictions, and has been influential in a broad range of areas (for example, genomics, climatology and advertising). The processes involved in the searching, curation and evaluation of data, and in quality control, are essential components of big-data practice, all of which have been the subject of conceptual exploration and formal methodological development in meta-analysis for many years<sup>64</sup>. However, the approach has been different from that taken for meta-analyses. Meta-analysis is inherently statistical, whereas big data has been framed within the field of computer science. Greater cross-disciplinary interactions should prove productive for both fields. Although formal systematic reviews and meta-analyses have long been established in many disciplines, they are only recently making inroads in fields such as molecular biology and genomics. Rapid gains in scientific progress stand to be made when these methods are more fully implemented throughout the biological sciences, and throughout science more generally.

Open-science practices have emphasized full and unbiased access to scientific data<sup>65</sup>, which is of longstanding importance and central to future progress in meta-analysis. Pre-registration (called 'registration' in some fields) of planned studies can reduce selective reporting of outcomes; publication of 'registered reports' in which the methods and proposed analyses for a study are peer-reviewed and published before the research is conducted can reduce publication bias. Limitations on accessing information are serious impediments for best practices in meta-analysis. By minimizing selective and poor reporting and advocating full access to the data and code associated with each analysis, open-science standards, including guidelines such as those in the Equator Network (<https://www.equator-network.org>)<sup>30,66</sup> can alleviate many problems in research synthesis and propel more rapid scientific advances.

In addition to the benefits that have been accrued from the increased availability of unbiased information, advances in meta-analytic techniques are being driven by methodological developments. Advances include: the use of machine learning and artificial intelligence (AI) to screen studies for inclusion in systematic reviews and meta-analyses<sup>67</sup>; increasingly sophisticated software and models for complex meta-regression<sup>17,47</sup>; robust variance estimation in studies with small sample sizes<sup>68</sup>; meta-analysis of individual participant data; and integration of meta-analysis and decision support in medicine and other fields<sup>69</sup>. Bayesian meta-analysis has been implemented in many fields and is a particularly useful approach when external sources of information can provide valid priors<sup>70</sup> or when a dataset is of sufficient quality and size that distributions can be fitted to it instead of attempting to fit it to familiar distributions. Meta-analytic approaches have been used to synthesize data to address methodological issues such as heterogeneity and its interpretation<sup>71</sup> and the implications of the inclusion or exclusion of unpublished literature<sup>72</sup>. Better integration of big data, AI and meta-analysis will depend on both conceptual

and methodological developments, and is reliant on greater trans-disciplinary links between statistics, computer science, the biological and social sciences, and other scientific fields. It is not impossible to envisage automated systems whereby AI aids not only in the real-time acquisition but also in the critical appraisal and meta-analysis of data, potentially integrating different information streams to inform tailored decisions in many areas of applied science.

The statistical methodologies that underpin and support meta-analysis have been undergoing continual development. Areas of particular current interest include multiple imputation to model missing data, advanced use of meta-regression and model selection to evaluate the influence of more complex data structures and multiple covariates, and hierarchical modelling of multi-level data, including that from individual 'participant' data in medicine<sup>22</sup> and in EEC<sup>73</sup>. Network meta-analyses seek to provide comparisons of multiple interventions, including indirect comparisons<sup>74</sup>. These methods are particularly useful when a set of randomized control trials with pairwise comparisons of interventions has been carried out with common interventions among the studies, but when not all studies include all interventions. Developments in and applications of this powerful approach have advanced considerably in clinical medicine over the past ten years<sup>75</sup>, providing better information about which treatment is most effective when there are multiple treatment options and pathways. 'Living' reviews, which are constantly updated, can prevent stale information from being cemented into belief or practice and have the potential to change the fundamental understanding of a problem or approach, because knowledge is being updated and new papers are being published continuously<sup>76</sup>. Rather than summarizing information in many individual reviews, living reviews and living cumulative network meta-analyses may also help to reduce waste in research by using the available primary studies more efficiently, by identifying gaps in research and by determining when the evidence is sufficient for decision and policy making<sup>77</sup>. However, their full implementation might require a reward shift both for primary researchers and synthesists.

Perhaps the most important foundation for advances in meta-analytic techniques is education in high-quality research-synthesis methods. Training in meta-analytic methods and concepts should be part of the basic training for higher-degree candidates in basic and applied scientific fields, including research post-graduates, medical doctors and other professional science practitioners (such as environmental consultants). This would formally embed their work in the context of existing evidence and facilitate learning of both statistical and critical appraisal skills. Those involved in primary research also need a better understanding of meta-analysis to exploit the revolution of open data fully. Most importantly, a new generation of scientists, peer reviewers, editors and science-policy practitioners would benefit from an increased understanding of the methodologies and interpretation of evidence synthesis.

Meta-analysis can be a key tool for facilitating rapid progress in science by quantifying what is known and identifying what is not yet known. Evidence synthesis should become a regular companion to primary scientific research to maximize the effectiveness of scientific inquiry. An evidence-based approach is important for progress in science, policy, and medical and conservation practice. This will require collaboration between statisticians, primary researchers and research synthesists, between meta-analysts and stakeholders, and among research synthesists across different disciplines. We are confident that, provided such collaborations are successful, meta-analysis will survive its 'midlife crisis' and emerge stronger and with a new-found purpose.

Received 4 March 2017; accepted 12 January 2018.

- Jennions, M. D., Lortie, C. J. & Koricheva, J. in *The Handbook of Meta-analysis in Ecology and Evolution* (eds Koricheva, J. et al.) Ch. 23, 364–380 (Princeton Univ. Press, 2013).
- Roberts, P. D., Stewart, G. B. & Pullin, A. S. Are review articles a reliable source of evidence to support conservation and environmental management? A comparison with medicine. *Biol. Conserv.* **132**, 409–423 (2006).
- Bastian, H., Glasziou, P. & Chalmers, I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med.* **7**, e1000326 (2010).

4. Borman, G. D. & Grigg, J. A. in *The Handbook of Research Synthesis and Meta-analysis* 2nd edn (eds Cooper, H. M. et al.) 497–519 (Russell Sage Foundation, 2009).
5. Ioannidis, J. P. A. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q.* **94**, 485–514 (2016).
6. Koricheva, J. & Gurevitch, J. Uses and misuses of meta-analysis in plant ecology. *J. Ecol.* **102**, 828–844 (2014).
7. Littell, J. H. & Shlonsky, A. Making sense of meta-analysis: a critique of “effectiveness of long-term psychodynamic psychotherapy”. *Clin. Soc. Work J.* **39**, 340–346 (2011).
8. Morrissey, M. B. Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *J. Evol. Biol.* **29**, 1882–1904 (2016).
9. Whittaker, R. J. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness-productivity relationship. *Ecology* **91**, 2522–2533 (2010).
10. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012); clarification **485**, 41 (2012).
11. Hillebrand, H. & Cardinale, B. J. A critique for meta-analyses and the productivity-diversity relationship. *Ecology* **91**, 2545–2549 (2010).
12. Moher, D. et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* **6**, e1000097 (2009).
- This paper provides a consensus regarding the reporting requirements for medical meta-analysis and has been highly influential in ensuring good reporting practice and standardizing language in evidence-based medicine, with further guidance for protocols, individual patient data meta-analyses and animal studies.**
13. Moher, D. et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst. Rev.* **4**, 1 (2015).
14. Nakagawa, S. & Santos, E. S. A. Methodological issues and advances in biological meta-analysis. *Evol. Ecol.* **26**, 1253–1274 (2012).
15. Nakagawa, S., Noble, D. W. A., Senior, A. M. & Lagisz, M. Meta-evaluation of meta-analysis: ten appraisal questions for biologists. *BMC Biol.* **15**, 18 (2017).
16. Hedges, L. & Olkin, I. *Statistical Methods for Meta-analysis* (Academic Press, 1985).
17. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 (2010).
18. Anzueto-Cabrera, J. & Higgins, J. P. T. Graphical displays for meta-analysis: an overview with suggestions for practice. *Res. Synth. Methods* **1**, 66–80 (2010).
19. Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *Br. Med. J.* **315**, 629–634 (1997).
20. Duval, S. & Tweedie, R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56**, 455–463 (2000).
21. Leimu, R. & Koricheva, J. Cumulative meta-analysis: a new tool for detection of temporal trends and publication bias in ecology. *Proc. R. Soc. Lond. B* **271**, 1961–1966 (2004).
22. Higgins, J. P. T. & Green, S. (eds) *Cochrane Handbook for Systematic Reviews of Interventions: Version 5.1.0* (Wiley, 2011).
- This large collaborative work provides definitive guidance for the production of systematic reviews in medicine and is of broad interest for methods development outside the medical field.**
23. Lau, J., Rothstein, H. R. & Stewart, G. B. in *The Handbook of Meta-analysis in Ecology and Evolution* (eds Koricheva, J. et al.) Ch. 25, 407–419 (Princeton Univ. Press, 2013).
24. Lortie, C. J., Stewart, G., Rothstein, H. & Lau, J. How to critically read ecological meta-analyses. *Res. Synth. Methods* **6**, 124–133 (2015).
25. Murad, M. H. & Montori, V. M. Synthesizing evidence: shifting the focus from individual studies to the body of evidence. *J. Am. Med. Assoc.* **309**, 2217–2218 (2013).
26. Rasmussen, S. A., Chu, S. Y., Kim, S. Y., Schmid, C. H. & Lau, J. Maternal obesity and risk of neural tube defects: a meta-analysis. *Am. J. Obstet. Gynecol.* **198**, 611–619 (2008).
27. Littell, J. H., Campbell, M., Green, S. & Toews, B. Multisystemic therapy for social, emotional, and behavioral problems in youth aged 10–17. *Cochrane Database Syst. Rev.* <https://doi.org/10.1002/14651858.CD004797.pub4> (2005).
28. Schmidt, F. L. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *Am. Psychol.* **47**, 1173–1181 (1992).
29. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013); erratum **14**, 451 (2013).
30. Parker, T. H. et al. Transparency in ecology and evolution: real problems, real solutions. *Trends Ecol. Evol.* **31**, 711–719 (2016).
31. Stewart, G. Meta-analysis in applied ecology. *Biol. Lett.* **6**, 78–81 (2010).
32. Sutherland, W. J., Pullin, A. S., Dolman, P. M. & Knight, T. M. The need for evidence-based conservation. *Trends Ecol. Evol.* **19**, 305–308 (2004).
33. Lowry, E. et al. Biological invasions: a field synopsis, systematic review, and database of the literature. *Ecol. Evol.* **3**, 182–196 (2013).
34. Parmesan, C. & Yohe, G. A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **421**, 37–42 (2003).
35. Jennions, M. D., Lortie, C. J. & Koricheva, J. in *The Handbook of Meta-analysis in Ecology and Evolution* (eds Koricheva, J. et al.) Ch. 24, 381–403 (Princeton Univ. Press, 2013).
36. Balvanera, P. et al. Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecol. Lett.* **9**, 1146–1156 (2006).
37. Cardinale, B. J. et al. Effects of biodiversity on the functioning of trophic groups and ecosystems. *Nature* **443**, 989–992 (2006).
38. Rey Benayas, J. M., Newton, A. C., Diaz, A. & Bullock, J. M. Enhancement of biodiversity and ecosystem services by ecological restoration: a meta-analysis. *Science* **325**, 1121–1124 (2009).
39. Leimu, R., Mutikainen, P. I. A., Koricheva, J. & Fischer, M. How general are positive relationships between plant population size, fitness and genetic variation? *J. Ecol.* **94**, 942–952 (2006).
40. Hillebrand, H. On the generality of the latitudinal diversity gradient. *Am. Nat.* **163**, 192–211 (2004).
41. Gurevitch, J. in *The Handbook of Meta-analysis in Ecology and Evolution* (eds Koricheva, J. et al.) Ch. 19, 313–320 (Princeton Univ. Press, 2013).
42. Rustad, L. et al. A meta-analysis of the response of soil respiration, net nitrogen mineralization, and aboveground plant growth to experimental ecosystem warming. *Oecologia* **126**, 543–562 (2001).
43. Adams, D. C. Phylogenetic meta-analysis. *Evolution* **62**, 567–572 (2008).
44. Hadfield, J. D. & Nakagawa, S. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J. Evol. Biol.* **23**, 494–508 (2010).
45. Lajeunesse, M. J. Meta-analysis and the comparative phylogenetic method. *Am. Nat.* **174**, 369–381 (2009).
46. Rosenberg, M. S., Adams, D. C. & Gurevitch, J. *MetaWin: Statistical Software for Meta-Analysis with Resampling Tests* Version 1 (Sinauer Associates, 1997).
47. Wallace, B. C. et al. OpenMEE: intuitive, open-source software for meta-analysis in ecology and evolutionary biology. *Methods Ecol. Evol.* **8**, 941–947 (2016).
48. Gurevitch, J., Morrison, J. A. & Hedges, L. V. The interaction between competition and predation: a meta-analysis of field experiments. *Am. Nat.* **155**, 435–453 (2000).
49. Adams, D. C., Gurevitch, J. & Rosenberg, M. S. Resampling tests for meta-analysis of ecological data. *Ecology* **78**, 1277–1283 (1997).
50. Gurevitch, J. & Hedges, L. V. Statistical issues in ecological meta-analyses. *Ecology* **80**, 1142–1149 (1999).
51. Schmid, C. H. & Mengersen, K. in *The Handbook of Meta-analysis in Ecology and Evolution* (eds Koricheva, J. et al.) Ch. 11, 145–173 (Princeton Univ. Press, 2013).
52. Eysenck, H. J. Exercise in mega-silliness. *Am. Psychol.* **33**, 517 (1978).
53. Simberloff, D. Rejoinder to: Don't calculate effect sizes; study ecological effects. *Ecol. Lett.* **9**, 921–922 (2006).
54. Cadotte, M. W., Mehrkens, L. R. & Menge, D. N. L. Gauging the impact of meta-analysis on ecology. *Evol. Ecol.* **26**, 1153–1167 (2012).
55. Koricheva, J., Jennions, M. D. & Lau, J. in *The Handbook of Meta-analysis in Ecology and Evolution* (eds Koricheva, J. et al.) Ch. 15, 237–254 (Princeton Univ. Press, 2013).
56. Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H. & Olkin, I. The case of the misleading funnel plot. *Br. Med. J.* **333**, 597–600 (2006).
57. Vetter, D., Rucker, G. & Storch, I. Meta-analysis: a need for well-defined usage in ecology and conservation biology. *Ecosphere* **4**, 1–24 (2013).
58. Mengersen, K., Jennions, M. D. & Schmid, C. H. in *The Handbook of Meta-analysis in Ecology and Evolution* (eds Koricheva, J. et al.) Ch. 16, 255–283 (Princeton Univ. Press, 2013).
59. Patsopoulos, N. A., Analatos, A. A. & Ioannidis, J. P. A. Relative citation impact of various study designs in the health sciences. *J. Am. Med. Assoc.* **293**, 2362–2366 (2005).
60. Kueffer, C. et al. Fame, glory and neglect in meta-analyses. *Trends Ecol. Evol.* **26**, 493–494 (2011).
61. Cohnstaedt, L. W. & Poland, J. Review Articles: The black-market of scientific currency. *Ann. Entomol. Soc. Am.* **110**, 90 (2017).
62. Longo, D. L. & Drazen, J. M. Data sharing. *N. Engl. J. Med.* **374**, 276–277 (2016).
63. Gauch, H. G. *Scientific Method in Practice* (Cambridge Univ. Press, 2003).
64. Science Staff. Dealing with data: introduction. Challenges and opportunities. *Science* **331**, 692–693 (2011).
65. Nosek, B. A. et al. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
66. Stewart, L. A. et al. Preferred reporting items for a systematic review and meta-analysis of individual participant data: the PRISMA-IPD statement. *J. Am. Med. Assoc.* **313**, 1657–1665 (2015).
67. Saldanha, I. J. et al. Evaluating Data Abstraction Assistant, a novel software application for data abstraction during systematic reviews: protocol for a randomized controlled trial. *Syst. Rev.* **5**, 196 (2016).
68. Tipton, E. & Pustejovsky, J. E. Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *J. Educ. Behav. Stat.* **40**, 604–634 (2015).
69. Mengersen, K., MacNeil, M. A. & Caley, M. J. The potential for meta-analysis to support decision analysis in ecology. *Res. Synth. Methods* **6**, 111–121 (2015).
70. Ashby, D. Bayesian statistics in medicine: a 25 year review. *Stat. Med.* **25**, 3589–3631 (2006).
71. Senior, A. M. et al. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. *Ecology* **97**, 3293–3299 (2016).
72. McAuley, L., Pham, B., Tugwell, P. & Moher, D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet* **356**, 1228–1231 (2000).
73. Koricheva, J., Gurevitch, J. & Mengersen, K. (eds) *The Handbook of Meta-Analysis in Ecology and Evolution* (Princeton Univ. Press, 2013).



**This book provides the first comprehensive guide to undertaking meta-analyses in ecology and evolution and is also relevant to other fields where heterogeneity is expected, incorporating explicit consideration of the different approaches used in different domains.**

74. Lumley, T. Network meta-analysis for indirect treatment comparisons. *Stat. Med.* **21**, 2313–2324 (2002).
75. Zarin, W. *et al.* Characteristics and knowledge synthesis approach for 456 network meta-analyses: a scoping review. *BMC Med.* **15**, 3 (2017).
76. Elliott, J. H. *et al.* Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Med.* **11**, e1001603 (2014).
77. Vandvik, P. O., Brignardello-Petersen, R. & Guyatt, G. H. Living cumulative network meta-analysis to reduce waste in research: a paradigmatic shift for systematic reviews? *BMC Med.* **14**, 59 (2016).
78. Jarvinen, A. A meta-analytic study of the effects of female age on laying date and clutch size in the Great Tit *Parus major* and the Pied Flycatcher *Ficedula hypoleuca*. *Ibis* **133**, 62–67 (1991).
79. Arnqvist, G. & Wooster, D. Meta-analysis: synthesizing research findings in ecology and evolution. *Trends Ecol. Evol.* **10**, 236–240 (1995).
80. Hedges, L. V., Gurevitch, J. & Curtis, P. S. The meta-analysis of response ratios in experimental ecology. *Ecology* **80**, 1150–1156 (1999).
81. Gurevitch, J., Curtis, P. S. & Jones, M. H. Meta-analysis in ecology. *Adv. Ecol. Res.* **32**, 199–247 (2001).
82. Lajeunesse, M. J. phyloMeta: a program for phylogenetic comparative analyses with meta-analysis. *Bioinformatics* **27**, 2603–2604 (2011).
83. Pearson, K. Report on certain enteric fever inoculation statistics. *Br. Med. J.* **2**, 1243–1246 (1904).
84. Fisher, R. A. *Statistical Methods for Research Workers* (Oliver and Boyd, 1925).
85. Yates, F. & Cochran, W. G. The analysis of groups of experiments. *J. Agric. Sci.* **28**, 556–580 (1938).
86. Cochran, W. G. The combination of estimates from different experiments. *Biometrics* **10**, 101–129 (1954).
87. Smith, M. L. & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *Am. Psychol.* **32**, 752–760 (1977).
88. Glass, G. V. Meta-analysis at middle age: a personal history. *Res. Synth. Methods* **6**, 221–231 (2015).
89. Cooper, H. M., Hedges, L. V. & Valentine, J. C. (eds) *The Handbook of Research Synthesis and Meta-analysis* 2nd edn (Russell Sage Foundation, 2009).
90. Rosenthal, R. *Meta-analytic Procedures for Social Research* (Sage, 1991).
91. Hunter, J. E., Schmidt, F. L. & Jackson, G. B. *Meta-analysis: Cumulating Research Findings Across Studies* (Sage, 1982).
92. Gurevitch, J., Morrow, L. L., Wallace, A. & Walsh, J. S. A meta-analysis of competition in field experiments. *Am. Nat.* **140**, 539–572 (1992).
93. O'Rourke, K. An historical perspective on meta-analysis: dealing quantitatively with varying study results. *J. R. Soc. Med.* **100**, 579–582 (2007).
94. Shadish, W. R. & Lecy, J. D. The meta-analytic big bang. *Res. Synth. Methods* **6**, 246–264 (2015).
95. Glass, G. V. Primary, secondary, and meta-analysis of research. *Educ. Res.* **5**, 3–8 (1976).
96. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177–188 (1986).
97. Lipsey, M. W. & Wilson, D. B. The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *Am. Psychol.* **48**, 1181–1209 (1993).
98. Chalmers, I. & Altman, D. G. *Systematic Reviews* (BMJ Publishing Group, 1995).
99. Moher, D. *et al.* Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. *Lancet* **354**, 1896–1900 (1999).
100. Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558 (2002).

**Acknowledgements** We dedicate this Review to the memory of Ingram Olkin and William Shadish, founding members of the Society for Research Synthesis Methodology who made tremendous contributions to the development of meta-analysis and research synthesis and to the supervision of generations of students. We thank L. Lagisz for help in preparing the figures. We are grateful to the Center for Open Science and the Laura and John Arnold Foundation for hosting and funding a workshop, which was the origination of this article. S.N. is supported by Australian Research Council Future Fellowship (FT130100268). J.G. acknowledges funding from the US National Science Foundation (ABI 1262402).

**Author Contributions** All authors contributed equally in designing the study and writing the manuscript, and so are listed alphabetically.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.G. ([jessica.gurevitch@stonybrook.edu](mailto:jessica.gurevitch@stonybrook.edu)), J.K. ([julia.koricheva@rhul.ac.uk](mailto:julia.koricheva@rhul.ac.uk)), S.N. ([s.nakagawa@unsw.edu.au](mailto:s.nakagawa@unsw.edu.au)) or G.S. ([gavin.stewart@newcastle.ac.uk](mailto:gavin.stewart@newcastle.ac.uk)).

**Reviewer Information** *Nature* thanks D. Altman, M. Lajeunesse, D. Moher and G. Romero for their contribution to the peer review of this work.

# GaN/NbN epitaxial semiconductor/superconductor heterostructures

Rusen Yan<sup>1\*</sup>, Guru Khalsa<sup>2\*</sup>, Suresh Vishwanath<sup>1</sup>, Yimo Han<sup>3</sup>, John Wright<sup>2</sup>, Sergei Rouvimov<sup>4</sup>, D. Scott Katzer<sup>5</sup>, Neeraj Nepal<sup>5</sup>, Brian P. Downey<sup>5</sup>, David A. Muller<sup>3,6</sup>, Huili G. Xing<sup>1,2,6</sup>, David J. Meyer<sup>5</sup> & Debdeep Jena<sup>1,2,6</sup>

**Epitaxy is a process by which a thin layer of one crystal is deposited in an ordered fashion onto a substrate crystal. The direct epitaxial growth of semiconductor heterostructures on top of crystalline superconductors has proved challenging. Here, however, we report the successful use of molecular beam epitaxy to grow and integrate niobium nitride (NbN)-based superconductors with the wide-bandgap family of semiconductors—silicon carbide, gallium nitride (GaN) and aluminium gallium nitride (AlGaIn). We apply molecular beam epitaxy to grow an AlGaIn/GaN quantum-well heterostructure directly on top of an ultrathin crystalline NbN superconductor. The resulting high-mobility, two-dimensional electron gas in the semiconductor exhibits quantum oscillations, and thus enables a semiconductor transistor—an electronic gain element—to be grown and fabricated directly on a crystalline superconductor. Using the epitaxial superconductor as the source load of the transistor, we observe in the transistor output characteristics a negative differential resistance—a feature often used in amplifiers and oscillators. Our demonstration of the direct epitaxial growth of high-quality semiconductor heterostructures and devices on crystalline nitride superconductors opens up the possibility of combining the macroscopic quantum effects of superconductors with the electronic, photonic and piezoelectric properties of the group III/nitride semiconductor family.**

The experimental discovery<sup>1</sup> of superconductivity in 1911 predated the controllable synthesis and understanding of semiconductors<sup>2</sup> by nearly three decades. However, in the time it took to uncover the correlated physics behind superconductivity, rapid advances in the band-theory of semiconductors, perfection in crystal growth, and discoveries such as donor- and acceptor-doping and quantum heterostructure<sup>3,4</sup> design had unleashed their technological potential, enabling electronic amplifiers and switches, as well as light-emitting diodes and diode lasers that operate at room temperature. These solid-state devices have replaced bulky and slow vacuum tubes and table-top lasers, and have shrunk information processing, storage, and communication systems onto a chip.

Today, semiconductor transistors are reaching their fundamental Boltzmann limits in terms of switching energy and power consumption in the digital von-Neumann computational architecture<sup>5</sup>, and communication systems are approaching their Shannon limits in terms of bandwidth and security. Quantum technologies have been envisaged to offer exponentially faster computation and guaranteed secure communications<sup>6</sup>, and the leading materials for these emerging technologies make use of the macroscopic manifestation of quantum properties in superconductors. Devices such as Josephson junction flux qubits<sup>7</sup>, lossless microwave resonators<sup>8</sup>, AC Josephson junction lasers<sup>9</sup> and superconducting single-photon detectors<sup>10</sup> are the building blocks of these new quantum-information systems.

Substantial advances in such systems would be expected if the power of semiconductors could be combined with that of superconductors on a single epitaxial platform<sup>11–13</sup>. The group III/nitride semiconductors GaN (with a bandgap,  $E_g$ , of about 3.4 eV), indium nitride (InN;  $E_g \approx 0.6$  eV) and AlN ( $E_g \approx 6.2$  eV) constitute the most revolutionary semiconductor family since silicon. That is because they offer, in a single heterostructure material family (see Fig. 1), the necessary

ingredients for ultrafast microwave communications<sup>14</sup>, ultralow-power computation<sup>15</sup>, high-voltage switches<sup>16</sup>, infrared through visible to deep-ultraviolet photonic emitters and detectors<sup>17,18</sup>, and high-frequency circuit components such as surface acoustic wave and bulk acoustic wave filters<sup>19</sup>. On the other hand, one of the most technologically important superconductor families comprises the nitride compounds NbN<sub>x</sub>, which have been used for superconducting radio-frequency circuits<sup>20</sup>, squid magnetometers<sup>21</sup>, Josephson junctions<sup>22</sup>, single-photon detectors<sup>10</sup> for quantum communications and astronomy, and a host of other applications<sup>23</sup>. Here, we report the successful epitaxial integration of the semiconducting and superconducting nitride families as a crucial enabler for several applications.

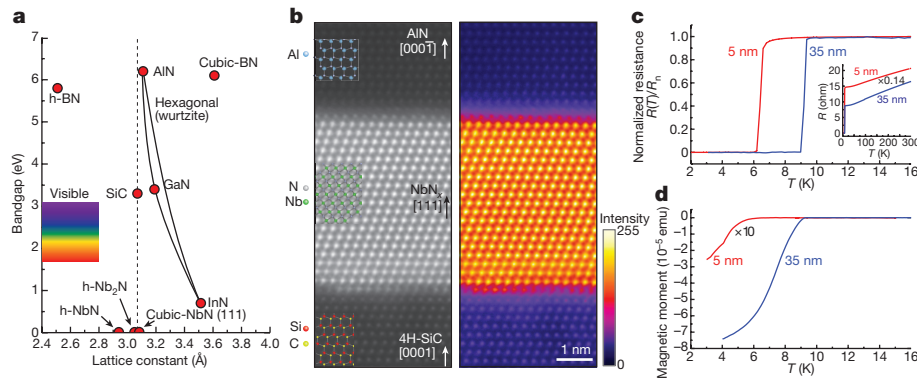
Figure 1a shows that the lattice constants of Nb-based nitride metals—such as hexagonal Nb<sub>2</sub>N and NbN, as well as cubic NbN rotated onto the (111) plane—are very close to the lattice constants of SiC, AlN and the GaN family. Wurtzite GaN and AlN can be grown on cubic (111) silicon, and hexagonal SiC serves as the substrate for the epitaxial growth of AlN- and GaN-based heterostructures for microwave transistors<sup>24</sup> and for quantum-well visible-light-emitting diodes<sup>18</sup>. Recently, we succeeded in growing crystalline epitaxial metal (epiMetal) niobium nitride layers by molecular beam epitaxy (MBE) on SiC, and further grew GaN and AlN layers on the epiMetal layers<sup>25,26</sup>. We found that the epiMetal layers retained high crystallinity and electronic conductivity down to thicknesses of a few nanometres<sup>25,26</sup>. The crystalline phases of the epilayers could be either hexagonal Nb<sub>2</sub>N or NbN, or cubic NbN. In this study, we have determined that our films are cubic NbN<sub>x</sub>, with  $x$  being around 0.75–0.88 as measured by secondary-ion mass spectrometry (SIMS). In what follows, we will simply refer to the phase and stoichiometry as NbN<sub>x</sub>. The use of NbN<sub>x</sub> enables an unprecedented level of epitaxial integration of buried metallic layers with wide-bandgap semiconductors and insulators.

<sup>1</sup>School of Electrical and Computer Engineering, Cornell University, Ithaca, New York 14853, USA. <sup>2</sup>Department of Materials Science and Engineering, Cornell University, Ithaca, New York 14853, USA. <sup>3</sup>School of Applied and Engineering Physics, Cornell University, Ithaca, New York 14853, USA. <sup>4</sup>Department of Electrical Engineering, University of Notre Dame, Indiana 46556, USA.

<sup>5</sup>Electronics Science and Technology Division, US Naval Research Laboratory, Washington DC 20375, USA. <sup>6</sup>Kavli Institute for Nanoscale Science, Cornell University, Ithaca, New York 14853, USA.

\*These authors contributed equally to this work.





**Figure 1 | Bandgap, lattice constant, crystallinity and superconductivity in epitaxial  $\text{NbN}_x$  on SiC.** **a**, Bandgap versus lattice constant for select nitride semiconductors as well as for SiC. **b**, Cross-section HAADF-STEM images in black/white (left) and false-colour (right) of 5-nm  $\text{NbN}_x$  grown on an SiC substrate with a AlN capping layer. **c**, Resistance versus temperature (normalized to the resistance at 16 K), showing the superconducting phase transition of 5-nm (red) and 35-nm (blue)

While investigating the low-temperature transport properties of the thin MBE-grown  $\text{NbN}_x$  layers, we find a superconducting phase transition at critical temperatures ( $T_c$ ) ranging from 6 K to 15 K, similar to what has been found for  $\text{NbN}_x$  grown by other methods<sup>27,28</sup>. Epitaxial layers of  $\text{NbN}_x$  thinner than the coherence length are found to exhibit two-dimensional superconductivity, with in-plane critical magnetic fields ( $H_c^{\parallel}$ ) well in excess of 20 T (the out-of-plane fields,  $H_c^{\perp}$ , are around 3 T).  $\text{NbN}_x$  is the first epitaxial superconductor to have been integrated with a technologically relevant semiconductor system.

### Growth of $\text{NbN}_x$ films by MBE

Niobium nitride used in superconducting electronics and bolometers for single-photon detectors, deposited by electron-beam evaporation or sputtering on non-epitaxial substrates, is typically polycrystalline<sup>10,21</sup>. Taking advantage of advances in MBE-based control of the growth of group III/nitride semiconductor heterostructures on SiC, we grew epitaxial layers of  $\text{NbN}_x$  directly on silicon-terminated, semi-insulating, four-hexagonal and six-hexagonal (4H and 6H) SiC substrates. We used a radio-frequency plasma nitrogen source of electronic-grade purity—identical to that used for AlN and GaN high-electron-mobility transistors (HEMTs), LEDs and lasers—to provide the active nitrogen atoms. We also used an electron-beam source of niobium, and monitored the growth *in situ* by reflection high-energy electron diffraction. Semiconducting Al(Ga)N/GaN quantum heterostructures were then grown epitaxially on top of the crystalline  $\text{NbN}_x$  layers.

Figure 1b shows high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) images of 5 nm  $\text{NbN}_x$  epitaxial layers grown on a semi-insulating 4H-SiC substrate and capped with an AlN layer. The epitaxial  $\text{NbN}_x$  layers are nearly completely cubic, with high crystalline quality over large areas. Occasional twin boundaries are seen—typically separated by about 1  $\mu\text{m}$ —as would be expected from the symmetry mismatch between cubic  $\text{NbN}_x$  and hexagonal SiC and AlN (see Extended Data Fig. 1). Figure 1b shows the epitaxial AlN on the  $\text{NbN}_x$  to be of nitrogen polarity; the entire AlN layer and all subsequent nitride semiconducting layers are hexagonal. The surfaces of uncapped  $\text{NbN}_x$  layers were extremely smooth, with a root-mean-square surface roughness of 0.16 nm for a 1  $\mu\text{m}$   $\times$  1  $\mu\text{m}$  region, as measured by atomic force microscopy (AFM; see Extended Data Fig. 2). Extended Data Fig. 3 shows X-ray diffraction (XRD) images of the epitaxial  $\text{NbN}_x$ .

### Electronic and magnetic properties of MBE-grown $\text{NbN}_x$

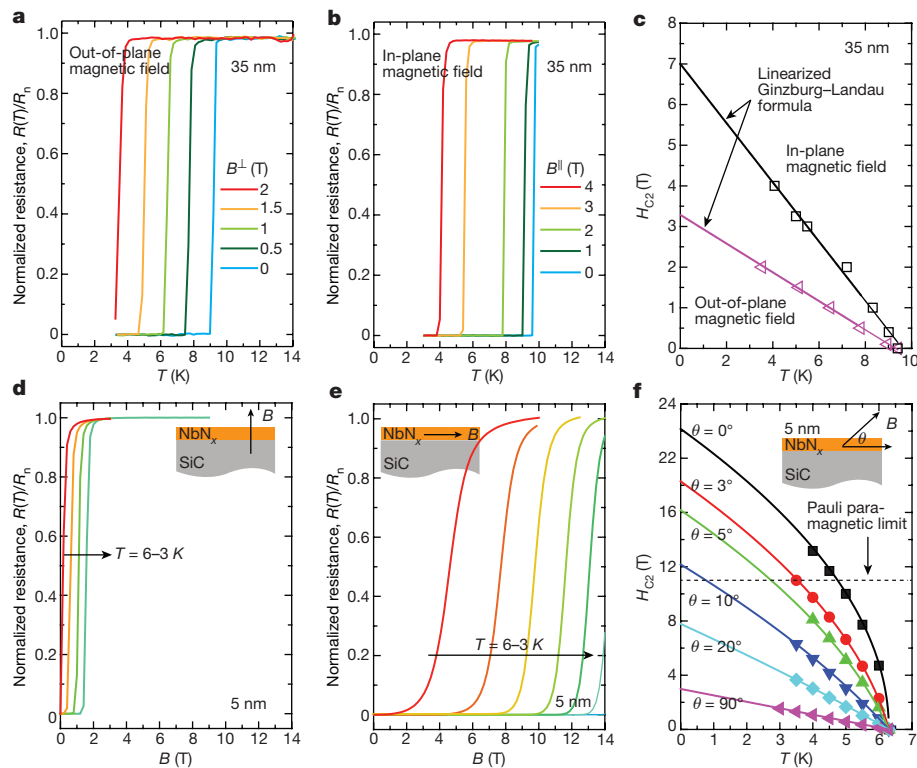
In its normal state we find that MBE  $\text{NbN}_x$  films are metallic with a resistivity of about  $10^{-5} \Omega \text{ cm}$ , comparable to that of bulk platinum

epitaxial  $\text{NbN}_x$  on SiC. Inset, resistance measured up to 300 K. **d**, The Meissner effect measured on the 5-nm and 35-nm samples, showing clear magnetic-flux expulsion accompanying the superconducting phase transition. These measurements are consistent with the  $T_c$  obtained in panel c.  $\times 10$  and  $\times 0.14$  indicate multiplication of the data by 10 or 0.14, respectively, to allow data of different scales to be shown on the same plot.

at room temperature. The measured Hall-effect carrier sign is negative, indicating electron conductivity, and the Hall-effect carrier density in three dimensions ( $n_{3d}$ ) is about  $2 \times 10^{23} \text{ cm}^{-3}$ , with a mean free path ( $\lambda$ ) of roughly  $1 - 2a_0$ , where  $a_0$  is the lattice constant (see Extended Data Table 1 for more metallic-state properties). Using a spherical Fermi surface approximation, the Mott–Ioffe–Regel criterion indicates that  $k_F\lambda$  is much greater than 1, where the Fermi wavevector ( $k_F$ ) is about  $(3\pi^2 n_{3d})^{1/3}$ , implying that the normal state transport is far above the minimum metallic conductivity regime. Although the Fermi surface is not spherical, we expect this conclusion to hold. We therefore find that our epitaxial  $\text{NbN}_x$  films are best characterized as working in the dirty limit ( $\lambda \ll \xi$ , where  $\xi$  is the coherence length), where the electron mean-free path is less than the Cooper-pair coherence length extracted from superconducting measurements, as described next.

Electrical transport measurements performed on the  $\text{NbN}_x$  layers, for thicknesses ranging from 4 nm to 100 nm, revealed superconductivity at transition temperatures of between 6 K and 15 K. Figure 1c shows the measured resistance  $R(T)$  normalized to the resistance at 16 K ( $R_n$ ) for  $\text{NbN}_x$  layers of thickness 5 nm and 35 nm. The resistivity of the samples exhibits a superconducting phase transition at around 7 K for the 5-nm sample, and about 9 K for the 35-nm sample. The inset shows the resistance up to 300 K for these two samples. In the metallic phase for temperatures  $T_c < T < 300 \text{ K}$ , the resistance shows an expected increase owing to phonon scattering. Figure 1d shows the Meissner effect measured on these two samples by vibrating sample magnetometry (VSM), revealing clear magnetic-flux expulsion accompanying the superconducting phase transition. The superconductivity transition temperature measured from electron transport and the Meissner effect are found to be consistent.

When the thickness of the semiconductor heterostructure quantum wells becomes smaller than the electron de-Broglie wavelength, quantum confinement drives signature two-dimensional effects such as the integer quantum Hall effect in single-particle magnetotransport<sup>29</sup>. Similarly, when the thickness of a superconducting layer  $d$  is less than the coherence length  $\xi$ , a high anisotropy in the Meissner effect upper critical field  $H_c^{\parallel}$  versus  $H_c^{\perp}$  is expected. These effects were recently reported in monolayer  $\text{NbSe}_2$ , a transition-metal dichalcogenide superconductor<sup>30</sup>. Figure 2a, b shows the out-of-plane and in-plane magnetic-field-dependent normalized resistance  $R(T)/R_n$  as a function of temperature for the 35-nm  $\text{NbN}_x$  epitaxial film. The variation of the critical field with the critical temperature is shown in Fig. 2c. Both out-of-plane and in-plane magnetic fields of strengths 0–4 T are seen to lower the critical temperature approximately linearly.



**Figure 2 | Magnetotransport measurements on 35-nm and 5-nm NbN<sub>x</sub> epitaxial films, showing two-dimensional superconductivity when the epilayer thickness is less than the coherence length. a–c, 35-nm NbN<sub>x</sub> films. d–f, 5-nm NbN<sub>x</sub> films. a, b, Temperature-dependent normalized resistance for the 35-nm sample, for various out-of-plane magnetic fields  $B^\perp$  (a) and in-plane fields  $B^\parallel$  (b). c, The critical field  $H_{c2}$  decreases linearly with temperature, consistent with the Ginzburg–Landau model of bulk superconductivity. d, For the 5-nm NbN<sub>x</sub> sample, the out-of-plane**

magnetic field destroys superconductivity easily at low fields. e, Much higher critical fields are needed when the field is in-plane. f, This strong anisotropy of critical fields is shown, plotting the critical field  $H_{c2}$  versus temperature for various angles,  $\theta$ , made by the magnetic field with the NbN<sub>x</sub> plane. The lines fit the linearized Ginzburg–Landau formula at  $\theta = 0^\circ$  and  $\theta = 90^\circ$ , and the lines for the intermediate angles are consistent with the Tinkham formula (see text).

The behaviour of the 5-nm-thick NbN<sub>x</sub> epitaxial layer is quite different. Figure 2d, e shows that substantially stronger in-plane magnetic fields compared with out-of-plane fields are required to break superconductivity for the 5-nm sample. The 5-nm sample remains superconducting at 3 K for in-plane fields up to 14 T, whereas the 35-nm sample is far into the metallic regime at this field.

The linearized Ginzburg–Landau equation for the perpendicular critical field is:

$$H_{c2}^\perp(T) = \frac{\phi_0}{2\pi\xi^2} \left( 1 - \frac{T}{T_c} \right) \quad (1)$$

where  $\phi_0 = h/2e$  is the superconducting flux quantum, with  $h$  being the Planck constant and  $2e$  the charge of a Cooper pair; and  $\xi = \xi_{GL}(0)$  is the extrapolation of the Ginzburg–Landau coherence length to  $T = 0$  K. From the  $\theta = 90^\circ$  fits in Fig. 2c, f, we extract  $\xi \approx 11$  nm for the  $d = 5$ -nm sample, and  $\xi \approx 10$  nm for the  $d = 35$ -nm sample. This explains our choice of representative sample thicknesses: one sample behaves like a thin film ( $d > \xi$ ) and one is in the two-dimensional limit ( $d < \xi$ ).

When the film thickness  $d$  is less than  $\xi$ , vortex formation under an in-plane magnetic field is severely suppressed. Because the density of Cooper pairs cannot change on a length scale shorter than  $\xi$ , vortices cannot accommodate flux for in-plane magnetic fields. Because for the  $d = 5$ -nm film  $d \leq \xi/2$ , Cooper-pair breaking caused by orbital effects requires a higher in-plane than out-of-plane magnetic field to destroy superconductivity. We believe that the Zeeman effect for pair-breaking<sup>31</sup> is suppressed in our NbN<sub>x</sub> films, and that orbital-pair-breaking is the dominant mechanism responsible for the

abnormally large  $H_{c2}^\parallel$  values. For in-plane critical fields, the Ginzburg–Landau formula in the two-dimensional limit is:

$$H_{c2}^\parallel(T) = \frac{\sqrt{12}\phi_0}{2\pi\xi d} \left( 1 - \frac{T}{T_c} \right)^{1/2} \quad (2)$$

With  $\xi$  extracted from equation (1), the effective superconducting thickness is extracted to be  $d = 4.9$  nm for the thin NbN<sub>x</sub> layer, in excellent agreement with the thickness measured by STEM. The extrapolation of this formula for  $\theta = 0^\circ$  in Fig. 2f to  $T \rightarrow 0$  K suggests an upper critical field  $H_{c2}^\parallel$  of about 22 T. This is twice the value of the Pauli paramagnetic limit,  $H_p$ , of about  $1.86 \times T_c$ —that is, 11 T—resulting from the Bardeen–Cooper–Schrieffer theory of superconductivity<sup>31</sup>. Such behaviour has also been observed in ultrathin superconducting systems: atomically thin layered transition-metal dichalcogenides<sup>30</sup>, ultrathin metals<sup>32</sup> and oxide heterojunctions<sup>33</sup> have all shown an anomalously large  $H_{c2}^\parallel$ . The possible reasons for this phenomenon are discussed further in the Methods.

We further ascertained the importance of the orbital-pair-breaking effect rather than the Zeeman effect by measuring the angle-dependent critical field for the thin NbN<sub>x</sub> sample. The results of angle-dependent magnetotransport measurements at  $\theta = 0^\circ, 3^\circ, 5^\circ, 10^\circ, 20^\circ$  and  $90^\circ$  for the 5-nm sample are shown in the  $H_{c2}$  versus  $T_c$  phase diagram in Fig. 2f;  $\theta$  is the angle that the magnetic-field vector makes with the NbN<sub>x</sub>/SiC heterointerface. The critical-field dependence on temperature changes from linear for  $\theta = 90^\circ$  to strongly nonlinear for  $\theta = 0^\circ$  for the 5-nm sample, whereas it remains linear for the 35-nm sample. As shown in Fig. 2f, the experimentally measured  $H_{c2}$  versus  $T_c$  at



intermediate angles at  $\theta = 3^\circ, 5^\circ, 10^\circ$  and  $20^\circ$  shows an exceptional agreement with the Tinkham formula<sup>34,35</sup>:

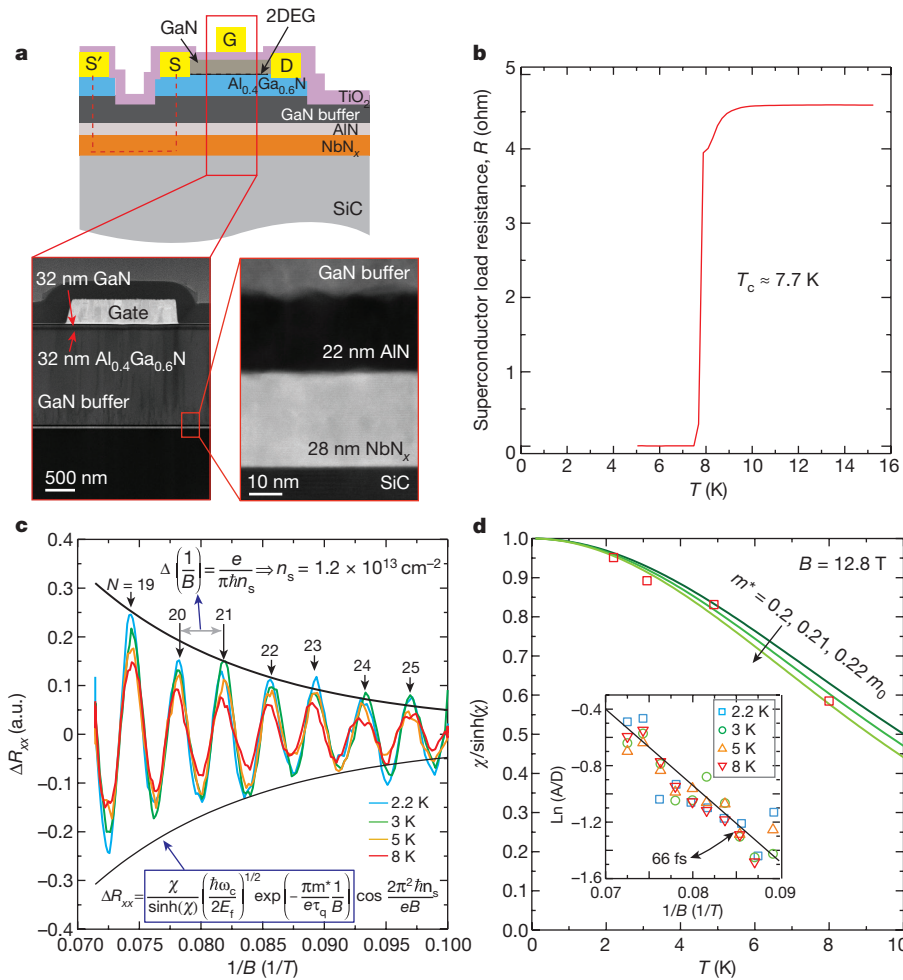
$$\left| \frac{H_{c2}(\theta, T) \sin \theta}{H_{c2}^\perp(T)} \right| + \left( \frac{H_{c2}(\theta, T) \cos \theta}{H_{c2}^\parallel(T)} \right)^2 = 1 \quad (3)$$

where  $H_{c2}^\perp(T)$  and  $H_{c2}^\parallel(T)$  are obtained from equations (1) and (2), and thus  $H_{c2}(\theta, T)$  is obtained by solving equation (3). Given that the Tinkham formula is obtained purely from the coupling between electron momentum and magnetic field<sup>34</sup>, the close agreement indicates that the observed pair-breaking is primarily a result of orbital effects, instead of the Zeeman effect (see Methods for further discussion). With the experimental determination of the critical temperature, coherence length and critical fields complete, we moved to the integration of nitride semiconductor heterostructures with epitaxial  $\text{NbN}_x$  films.

### Semiconductor/superconductor heterojunctions

The ability to grow epitaxial  $\text{Al}(\text{Ga})\text{N}$  and  $\text{GaN}$  on  $\text{NbN}_x$  has created an opportunity for the intimate integration of semiconductors with

superconductors. To demonstrate this functionality, we have grown a  $\text{GaN}/\text{AlGa}\text{N}$  quantum-well heterostructure on the buried epitaxial  $\text{NbN}_x$  superconducting layer, as shown in Fig. 3a. After epitaxial growth of 28-nm  $\text{NbN}_x$  on SiC, a 22-nm  $\text{AlN}$  layer, a 1.3- $\mu\text{m}$   $\text{GaN}$  buffer layer, a 32-nm  $\text{Al}_{0.4}\text{Ga}_{0.6}\text{N}$  barrier, and a 32-nm  $\text{GaN}$  channel layer are grown successively by MBE in a single run without breaking vacuum. The entire  $\text{AlN}/\text{GaN}/\text{AlGa}\text{N}/\text{GaN}$  heterostructure takes a nitrogen-polar wurtzite form of high crystallinity and has a sharp heterojunction. This is confirmed by Hall-effect measurements of the mobility ( $\mu$ ) of a two-dimensional electron gas (2DEG) of about  $1,350 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  at 300 K and about  $3,400 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  at 2 K, with two-dimensional densities ( $n_{2d}$ ) of about  $1.3 \times 10^{13} \text{ cm}^{-2}$  at 300 K and  $1.2 \times 10^{13} \text{ cm}^{-2}$  at 2 K. The 2DEG is formed in a triangular quantum well that is produced at the top  $\text{GaN}/\text{Al}_{0.4}\text{Ga}_{0.6}\text{N}$  heterojunction owing to the Berry-phase-driven spontaneous and piezoelectric polarization difference between  $\text{AlGa}\text{N}$  and  $\text{GaN}$ <sup>36</sup>. The high 2DEG mobility is comparable to that obtained in similar heterostructures without the  $\text{NbN}_x$  buried layer, indicating a successful epitaxial integration. The Hall-effect measurement also proves that the 2DEG is electrically isolated from the buried  $\text{NbN}_x$  metal layer. This 2DEG channel has enabled the integration of



**Figure 3 | Electrical and magnetotransport characterizations of group III/nitride/ $\text{NbN}_x$  heterostructures.** **a**, Cross-section schematic (top) and scanning transmission electron microscopy (STEM) imaging (bottom left and right) of  $\text{Al}(\text{Ga})\text{N}/\text{GaN}$  HEMTs/ $\text{NbN}_x$  grown by MBE on SiC substrates. **b**, Four-probe resistance of the buried epitaxial  $\text{NbN}_x$  layer, showing that it remains superconducting—with a  $T_c$  of about 7.7 K—after the subsequent growth of the HEMT. **c**, Measured  $\Delta R_{xx}$  versus  $1/B$  at various temperatures, extracted from the longitudinal resistivity ( $R_{xx}$ ) versus  $1/B$  after background subtraction (see Methods). The resistance oscillation period,  $\Delta(1/B)$ , is  $0.0038 \text{ T}^{-1}$ , which can be used to estimate

carrier concentration as  $n_s = 1.26 \times 10^{13} \text{ cm}^{-2}$ . The numbers on the arrows indicate the Landau level indices. **d**,  $\chi/\sinh(\chi)$  as a function of temperature at  $B = 12.8 \text{ T}$ . The lines are fittings made using effective masses,  $m^*$ , of  $0.2 m_e$ ,  $0.21 m_e$  and  $0.22 m_e$ . The inset shows Dingle plots at various temperatures, allowing extraction of the quantum-scattering time  $\tau_q$ . The linear fit to experimental data gives  $\tau_q = 66 \text{ fs}$ , which translates to a momentum/quantum-scattering ratio of  $\tau_i/\tau_q = 5.6 \gg 1$ —a clear indication of charged dislocations as the dominant scattering mechanism in this 2DEG<sup>42</sup>.

an HEMT with NbN<sub>x</sub>; before describing this integration, we discuss the quantum-transport properties of the 2DEG channel as probed by low-temperature magnetoresistance.

Low-temperature and high-magnetic-field measurements revealed clear Shubnikov–de Haas oscillations in the magnetoresistance of the 2DEG (Fig. 3c, d). These oscillations are commensurate with the magnetic-field-driven formation of Landau levels, and are used to extract the carrier concentration, electron effective mass, and quantum-scattering times<sup>37,38</sup> by using the Lifshitz–Kosevich<sup>39</sup> form of the magnetoresistance:

$$\Delta\rho \propto \mathcal{R}_T \mathcal{R}_D \cos\left(\frac{2\pi^2 \hbar n_{\text{sdH}}}{eB}\right)$$

In this equation, the periodicity in inverse magnetic field depends only on the carrier concentration  $n_{\text{sdH}}$  and the fundamental constants  $e$  and  $\hbar$ . The measured period of  $\Delta(1/B) = 0.0038 \text{ T}^{-1}$  shown in Fig. 3c corresponds to a carrier concentration of  $1.26 \times 10^{13} \text{ cm}^{-2}$ , consistent with low-field Hall-effect measurements.  $\mathcal{R}_T = \chi/\sinh(\chi)$  measures the thermal damping owing to a broadening of Landau levels, with the dimensionless factor  $\chi = 2\pi^2 k_B T/\hbar\omega_c$  parametrizing the ratio of the thermal energy to the Landau-level energy separation<sup>39</sup>. Here  $k_B$  is the Boltzmann constant,  $T$  is the temperature, and  $\omega_c = eB/m^*$  is the cyclotron frequency with effective mass  $m^*$ . Figure 3d shows the factor  $\mathcal{R}_T$  plotted against the temperature dependence of the  $N = 19$  Landau-level peak amplitude. The effective mass  $m^* \approx 0.21m_e$  extracted from this plot is consistent with prior reports for 2DEGs in GaN<sup>40</sup>. Using the measured effective mass, the Dingle factor  $\mathcal{R}_D = \pi m^*/e\tau_q B$  reveals the quantum-scattering lifetime  $\tau_q$  (ref. 41). The inset of Fig. 3d shows that the peak amplitude varies with inverse magnetic field for various temperatures with a characteristic quantum-scattering time of about 66 fs. This value is substantially smaller than the transport-scattering time ( $\tau_t$ ) extracted from the low-temperature Hall-mobility measurement; the ratio  $\tau_t/\tau_q = 5.6$ , being much greater than 1, suggests that Coulomb scattering from charged dislocations is the dominant scattering mechanism in the 2DEG<sup>42</sup>. Dislocations of density of about  $10^9 \text{ cm}^{-2}$  are typically present in GaN/AlGaIn 2DEGs grown on SiC, Si or other substrates<sup>42–44</sup>. We emphasize that the presence of magnetic quantum oscillations demonstrates the high-quality epitaxial growth of the GaN/AlGaIn 2DEG on the superconducting NbN<sub>x</sub> film.

We fabricated nitrogen-polar GaN HEMTs as described in ref. 26. Low-resistance source/drain ohmic contacts were formed to the polarization-induced 2DEG, and 10 nm TiO<sub>2</sub> high-K dielectric was used before depositing the gate metal. Details of the process and device dimensions are described in the Methods. To form an electrical contact to the NbN<sub>x</sub> layer, we applied a large voltage between two adjacent metal contacts, S and S', that were initially isolated from each other by mesa etching (Fig. 3a). This process formed a low-resistance contact between S and S' through the epitaxial NbN<sub>x</sub> layer (dashed red line in Fig. 3a). A four-probe resistance measurement on such contacts (see Fig. 3b and Methods) confirmed that the buried NbN<sub>x</sub> epilayer retained its superconductivity, with a transition temperature of around 7.7 K, even after the epitaxial growth of the entire nitride heterostructure on top of it and the subsequent device processing and annealing steps.

Figure 4a shows the HEMT drain current ( $J_d$ ) per unit width,  $J_d = I_d/W$ , in logarithmic scale as a function of the gate voltage for two drain voltages at 5 K. Note that the gate voltage  $V_{gs'}$  (the voltage difference between gate  $g$  and source  $s'$ ) and drain voltage  $V_{ds'}$  (the voltage difference between drain  $d$  and source  $s'$ ) are measured with the buried NbN<sub>x</sub> layer serving as the source load of the HEMT. The gate leakage current is low, and the drain current changes by about six to seven orders of magnitude as the Fermi level of the GaN quantum-well channel is pulled from inside the conduction band at  $V_{gs'} = 0 \text{ V}$  into the gap at  $V_{gs'} = -8 \text{ V}$ . The high on/off ratio was also observed at room temperature, as shown in Extended Data Fig. 4 and discussed in the Methods.

To quantify the effect of the superconducting load element, we compare the current in the HEMT from the drain ( $D$ ) to S and S' under varying gate voltages. The  $J_d$ – $V_{gs'}$  transfer curve measured at 5 K deviates from the  $J_d$ – $V_{gs'}$  transfer curve for currents of greater than  $0.1 \text{ A mm}^{-1}$  (Extended Data Fig. 5). Below  $0.1 \text{ A mm}^{-1}$ , NbN<sub>x</sub> remains superconducting with  $R_{sc} = 0 \Omega$ , and therefore does not contribute to the measured transfer curve. A current larger than  $0.1 \text{ A mm}^{-1}$  drives the NbN<sub>x</sub> into a normal metal state with  $R_{sc} \approx 4.6 \Omega$ . The superconductor-to-metal phase transition can occur when the magnetic field is greater than the critical magnetic field ( $H_c$ ), when the current density is higher than the critical current density ( $J_c$ ), or when the temperature is higher than the critical temperature ( $T_c$ ). According to Ampere's law, the magnetic field resulting from the 2DEG current  $J_d$  at the superconducting layer is around  $\mu_0 J_d/2 \approx 10^{-4} \text{ T}$  (that is, much less than  $H_c$ ). As shown in Extended Data Fig. 6, we have measured the critical current density of the MBE NbN<sub>x</sub> to be  $J_c = 10^5 \text{ A cm}^{-2}$ , and for a thickness of  $t = 28 \text{ nm}$  the net current density is estimated to be  $J = 10^3 \text{ A cm}^{-2}$  (much less than  $J_c$ ). Thus, we rule out the Meissner effect and high current injection as possible causes of the superconductor-to-metal transition driven by the transistor. We attribute the transition to Joule heating at the semiconductor/superconductor junction. The abrupt appearance of a resistive load lowers the measured transistor current flowing across  $D$ -to- $S'$ , changing the transfer curve. Further investigation of this electronic phase change in the load shows that the effect is strong enough to drive a negative differential resistance (NDR) in the transistor output characteristics.

Figure 4b–d shows the measured  $J_d$ – $V_{ds'}$  output characteristics of the HEMT as a function of gate voltages, measured at 10 K, 7 K and 5 K, with the NbN<sub>x</sub> layer as the source load. At temperatures of 10 K (greater than  $T_c$ ), the NbN<sub>x</sub> layer acts as a resistive load at all bias conditions, and  $J_d$  increases monotonically with  $V_{ds'}$  (Fig. 4b). As the temperature is lowered to 7 K (less than  $T_c$ ), the NbN<sub>x</sub> load drops to its zero-resistance state. This is characterized by a lower transistor on-resistance and a weak NDR (Fig. 4c).

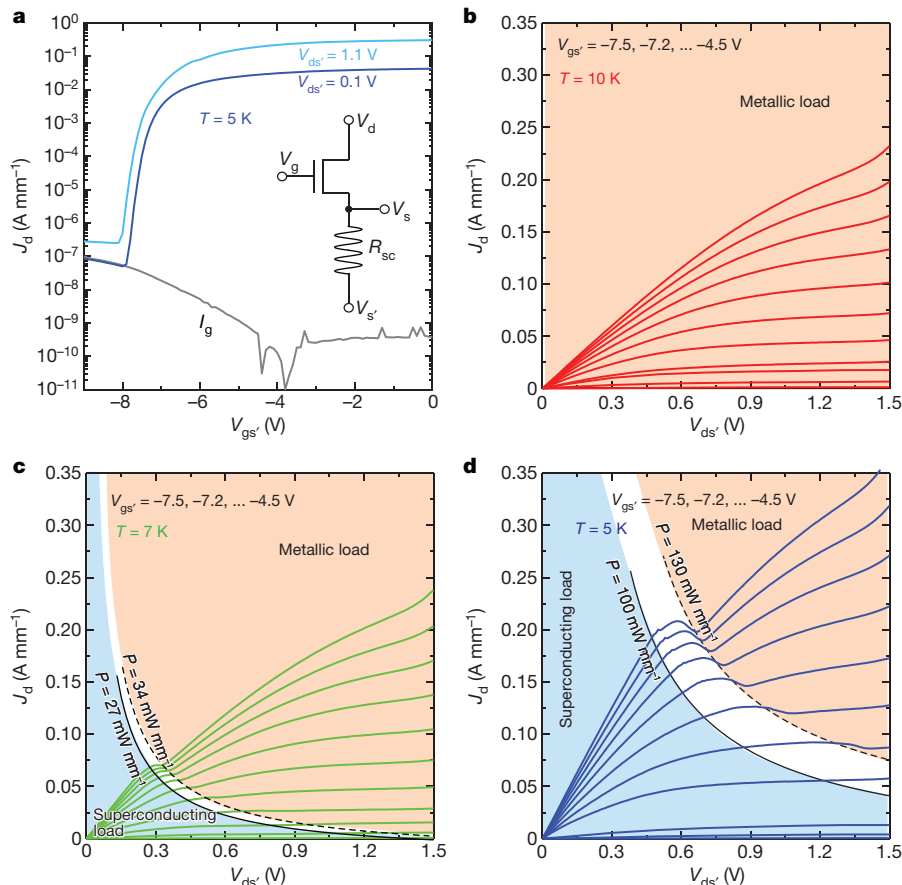
As the power level is increased, Joule heating warms the NbN<sub>x</sub>/AlN/GaN junction to temperatures higher than  $T_c$ , turning the surrounding superconducting NbN<sub>x</sub> into a normal metal, and thus lowering the channel current. The abrupt increase in resistance caused by the superconductor-to-metal transition leads directly to the appearance of an NDR, as seen clearly seen in Fig. 4c, d. The transition regime of load from superconducting phase to normal metal for all  $J_d$ – $V_{ds'}$  curves at 7 K and 5 K lies within two iso-power contours,  $P = I \times V$  (solid and dashed lines in Fig. 4c, d). This is a clear indication that the phase transition is thermally induced by Joule heating, and not through the critical current or critical magnetic field of NbN<sub>x</sub>. A critical-current-mediated manifestation or a critical-magnetic-field-mediated phase transition would have caused an NDR at the same current level, not the same power level.

This form of a phase-transition element attached at the source contact of a transistor has been used to demonstrate sub-Boltzmann switching in silicon and GaN transistors at room temperature<sup>45</sup>. In such phase-field-effect transistors, the phase change was obtained through a filamentary metal-to-insulator transition in VO<sub>2</sub> that was driven through a combination of thermal phase transition and Mott–Hubbard interactions by injected current. The superconducting phase transition at a low temperature in the hybrid superconductor–transistor phase-field-effect transistors and the resulting NDR behaviour has not been observed before.

## Conclusions

The successful epitaxial integration of group III/nitride semiconductors and transistor gain elements with NbN<sub>x</sub>-based superconductors points towards several new opportunities. Just as the development of reduced surface and interface states of silicon paved the way for the metal-oxide-semiconductor field-effect transistor, so do epitaxial NbN<sub>x</sub>/group III/nitride structures offer the possibility of defect-free metal/semiconductor heterojunctions. Semiconductor transistors were





**Figure 4 | Current–voltage characterizations of HEMTs with a superconducting source load at low temperatures.** **a**, Drain current density versus gate–source voltage ( $J_d$ – $V_{gs}$ ) transfer curves of the HEMTs at 5 K, showing a high on/off ratio at  $V_{ds} = 0.1$  V and 1.1 V. The inset shows the equivalent circuit diagram for the device. **b–d**,  $J_d$  versus source–drain voltage,  $V_{ds}$ , for various top–gate voltages,  $V_{gs}$ , of GaN HEMTs with

a buried epitaxial superconductor load at the source side, at temperatures of 10 K (**b**), 7 K (**c**), and 5 K (**d**). The results show that when the  $\text{NbN}_x$  layer becomes superconducting, the transistor output characteristics exhibit a negative differential resistance (NDR), as seen by the decrease in resistance with increasing  $V_{ds}$  between the iso-power contours. The black solid and dashed lines in panels **c**, **d** indicate iso-power contours.

instrumental in the discovery of the quantum-Hall effect<sup>29</sup>, which led to the discovery of topological insulators<sup>46</sup> and introduced topology into condensed-matter physics<sup>47</sup>. Epitaxial integration of semiconductor/superconductor heterostructures could enable phenomena that require both materials families, such as the Majorana zero-modes for braiding-based, topologically protected quantum computation<sup>11,48</sup>. Moreover, the presence of spontaneous and piezoelectric polarization induced by broken inversion symmetry in group III/nitride semiconductor crystals<sup>49</sup> offers the possibility of Rashba-driven topological insulators<sup>50</sup>. In more near-term applications,  $\text{NbN}_x$ -based single-photon detectors can now be epitaxially integrated with GaN HEMT amplifiers for secure quantum communications. Finally, combining GaN HEMT microwave amplifiers with  $\text{NbN}_x$ -based Josephson junctions can provide an all-epitaxial platform for superconducting qubits whereby the most desirable properties of semiconductors and superconductors are combined epitaxially in a seamless braid.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 July 2017; accepted 16 January 2018.

- Onnes, H. K. *Investigations into the Properties of Substances at Low Temperatures, Which Have Led, Amongst Other Things, to the Preparation of Liquid Helium.* (Nobel Lectures, 1913).
- Riordan, M. & Hodgeson, L. in *Crystal Fire* 88–90 (WW Norton and Company, 1998).
- Kroemer, H. Nobel lecture. Quasielectric fields and band offsets: teaching electrons new tricks. *Rev. Mod. Phys.* **73**, 783–793 (2001).

- Alferov, Z. I. Nobel lecture. The double heterostructure concept and its applications in physics, electronics, and technology. *Rev. Mod. Phys.* **73**, 767–782 (2001).
- Jena, D. Tunneling transistors based on graphene and 2-D crystals. *Proc. IEEE* **101**, 1585–1602 (2013).
- Ladd, T. D. *et al.* Quantum computers. *Nature* **464**, 45–53 (2010).
- Mooij, J. *et al.* Josephson persistent-current qubit. *Science* **285**, 1036–1039 (1999).
- Lancaster, M. *et al.* Superconducting microwave resonators. In *IEEE Proceedings H (Microwaves, Antennas and Propagation)*, vol. 139, 149–156 (IET, 1992).
- Cassidy, M. C. *et al.* Demonstration of an ac Josephson junction laser. *Science* **355**, 939–942 (2017).
- Gol'tsman, G. N., Okunev, O., Lipatov, A., Semenov, K., Voronov, B. & Dzardanov, A. Picosecond superconducting single-photon optical detector. *Appl. Phys. Lett.* **79**, 705–707 (2001).
- Mourik, V. *et al.* Signatures of Majorana fermions in hybrid superconductor-semiconductor nanowire devices. *Science* **336**, 1003–1007 (2012).
- Sarma, S. D., Freedman, M. & Nayak, C. Majorana zero modes and topological quantum computation. *npj Quant. Information* **1**, 15001 (2015).
- Krogstrup, P. *et al.* Epitaxy of semiconductor–superconductor nanowires. *Nat. Mater.* **14**, 400–406 (2015).
- Yue, Y. *et al.* Ultrascaled InAlN/GaN high electron mobility transistors with cutoff frequency of 400 GHz. *Jpn. J. Appl. Phys.* **52**, 08JN14 (2013).
- Li, W. *et al.* Polarization-engineered III-nitride heterojunction tunnel field-effect transistors. Exploratory solid-state computational devices and circuits. *IEEE J. Exp. Solid State Comp. Devices Circuits* **1**, 28–34 (2015).
- Hu, Z. *et al.* Near unity ideality factor and Shockley-Read-Hall lifetime in GaN-on-GaN pn diodes with avalanche breakdown. *Appl. Phys. Lett.* **107**, 243501 (2015).
- Islam, S. M. *et al.* MBE-grown 232–270 nm deep-UV LEDs using monolayer thin binary GaN/AlN quantum heterostructures. *Appl. Phys. Lett.* **110**, 041108 (2017).
- Sheu, J.-K. *et al.* White-light emission from near UV InGaIn-GaN LED chip precoated with blue/green/red phosphors. *IEEE Photonics Technol. Lett.* **15**, 18–20 (2003).

19. Dubois, M.-A. & Muller, C. in *MEMS-based Circuits and Systems for Wireless Communication* (eds Enz, C. C. & Kaiser, A.) 3–28 (Springer, 2013).
20. Pernice, W. H. *et al.* High-speed and high-efficiency travelling wave single-photon detectors embedded in nanophotonic circuits. *Nat. Commun.* **3**, 1325 (2012).
21. Faucher, M. *et al.* Niobium and niobium nitride SQUIDs based on anodized nanobridges made with an atomic force microscope. *Physica C* **368**, 211–217 (2002).
22. Song, S., Jin, B., Yang, H., Ketterson, J. & Schuller, I. K. Preparation of large area NbN/AlN/NbN Josephson junctions. *Jpn. J. Appl. Phys.* **26**, 1615 (1987).
23. Hajenius, M. *et al.* Low noise NbN superconducting hot electron bolometer mixers at 1.9 and 2.5 THz. *Supercond. Sci. Technol.* **17**, S224 (2004).
24. Eastman, L. F. & Mishra, U. K. The toughest transistor yet. *IEEE Spectr.* **39**, 28 (2002).
25. Katzer, D. S. *et al.* Epitaxial metallic  $\beta$ -Nb<sub>2</sub>N films grown by MBE on hexagonal SiC substrates. *Appl. Phys. Exp.* **8**, 085501 (2015).
26. Meyer, D. J. *et al.* Epitaxial lift-off and transfer of III-N materials and devices from SiC substrates. *IEEE Trans. Semicond. Manuf.* **29**, 384–389 (2016).
27. Sanjinés, R., Benkahoul, M., Sandu, C., Schmid, P. & Lévy, F. Electronic states and physical properties of hexagonal  $\beta$ -Nb<sub>2</sub>N and  $\delta'$ -NbN nitrides. *Thin Solid Films* **494**, 190–195 (2006).
28. Meyer, D. J. *et al.* N-polar n+ GaN cap development for low ohmic contact resistance to inverted HEMTs. *Phys. Status Solidi C* **9**, 894–897 (2012).
29. Klitzing, K. v., Dorda, G. & Pepper, M. New method for high-accuracy determination of the fine-structure constant based on quantized Hall resistance. *Phys. Rev. Lett.* **45**, 494 (1980).
30. Xi, X. *et al.* Ising pairing in superconducting NbSe<sub>2</sub> atomic layers. *Nat. Phys.* **12**, 139–143 (2016).
31. Clogston, A. M. Upper limit for the critical field in hard superconductors. *Phys. Rev. Lett.* **9**, 266 (1962).
32. Nam, H. *et al.* Ultrathin two-dimensional superconductivity with strong spin–orbit coupling. *Proc. Natl Acad. Sci. USA* **113**, 10513–10517 (2016).
33. Kozuka, Y. *et al.* Two-dimensional normal-state quantum oscillations in a superconducting heterostructure. *Nature* **462**, 487–490 (2009).
34. Tinkham, M. Effect of fluxoid quantization on transitions of superconducting films. *Phys. Rev.* **129**, 2413 (1963).
35. Aoi, K., Meserve, R. & Tedrow, P. Hc (0) and Tinkham's formula for high-field superconductors. *Phys. Rev. B* **7**, 554 (1973).
36. Ambacher, O. *et al.* Two-dimensional electron gases induced by spontaneous and piezoelectric polarization charges in N- and Ga-face AlGaIn/GaN heterostructures. *J. Appl. Phys.* **85**, 3222–3233 (1999).
37. Jena, D. *et al.* Magnetotransport properties of a polarization-doped three-dimensional electron slab. *Phys. Rev. B* **67**, 153306 (2003).
38. Cao, Y., Wang, K., Orlov, A., Xing, H. & Jena, D. Very low sheet resistance and Shubnikov-de Haas oscillations in two-dimensional electron gases at ultrathin binary AlN/GaN heterojunctions. *Appl. Phys. Lett.* **92**, 152112 (2008).
39. Hamaguchi, C. *Basic Semiconductor Physics* (Springer, 2001).
40. Manfra, M. J. *et al.* Electron mobility exceeding 160 000 cm<sup>2</sup>/Vs in AlGaIn/GaN heterostructures grown by molecular-beam epitaxy. *Appl. Phys. Lett.* **85**, 5394–5396 (2004).
41. Dingle, R. Some magnetic properties of metals. II. The influence of collisions on the magnetic behaviour of large systems. *Proc. R. Soc. Lond. A* **211**, 517–525 (1952).
42. Jena, D. & Mishra, U. K. Quantum and classical scattering times due to charged dislocations in an impure electron gas. *Phys. Rev. B* **66**, 241307 (2002).
43. Hsu, J. *et al.* Effect of growth stoichiometry on the electrical activity of screw dislocations in GaN films grown by molecular-beam epitaxy. *Appl. Phys. Lett.* **78**, 3980–3982 (2001).
44. Kaun, S. W., Wong, M. H., Mishra, U. K. & Speck, J. S. Correlation between threading dislocation density and sheet resistance of AlGaIn/AlN/GaN heterostructures grown by plasma-assisted molecular beam epitaxy. *Appl. Phys. Lett.* **100**, 262102 (2012).
45. Shukla, N. *et al.* A steep-slope transistor based on abrupt electronic phase transition. *Nat. Commun.* **6**, 7812 (2015).
46. Thouless, D. J., Kohmoto, M., Nightingale, M. P. & den Nijs, M. Quantized Hall conductance in a two-dimensional periodic potential. *Phys. Rev. Lett.* **49**, 405 (1982).
47. Hasan, M. Z. & Kane, C. L. Colloquium. Topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
48. Beenakker, C. Search for Majorana Fermions in superconductors. *Annu. Rev. Condens. Matter Phys.* **4**, 113–136 (2013).
49. Wood, C. & Jena, D. *Polarization Effects in Semiconductors: From Ab-Initio Theory to Device Applications* (Springer, 2007).
50. Miao, M. S. *et al.* Polarization-driven topological insulator transition in a GaN/InN/GaN quantum well. *Phys. Rev. Lett.* **109**, 186803 (2012).

**Acknowledgements** We thank A.H. MacDonald for fruitful discussions, and D. Storm for facilitating SIMS measurements. For the measurements performed here, we made use of the Cornell Center for Materials Research (CCMR) Shared Facilities, which are supported through the National Science Foundation (NSF) Materials Research Science and Engineering Centers (MRSEC) program (grant DMR-1719875). The structure fabrications were realized in part at the Cornell NanoScale Facility, a member of the National Nanotechnology Coordinated Infrastructure (NNCI), which is supported by the NSF (grant ECCS-1542081), and a CCMR Superconductor Seed. D.J. and D.J.M. acknowledge funding support from the Office of Naval Research, monitored by P. Maki. D.J.M. also acknowledges device processing support from N. Green.

**Author Contributions** R.Y., S.V. and J.W. performed electrical, magnetic and magnetotransport measurements. D.S.K., N.N., B.P.D. and D.J.M. grew and characterized the epitaxial layers. Y.H. performed scanning transmission electron microscopy (STEM) analysis on thin NbN, films under the supervision of D.A.M. S.R. conducted the transmission electron microscopy (TEM) measurements. R.Y. and G.K. conducted experimental data analysis and theoretical calculations, with help from D.J. and H.G.X. R.Y., G.K. and D.J. wrote the manuscript, with input from all authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to R.Y. (ry253@cornell.edu), D.J.M. (david.meyer@nrl.navy.mil) or D.J. (djena@cornell.edu).

**Reviewer Information** Nature thanks Y. Krockenberger and the other anonymous reviewer(s) for their contribution to the peer review of this work.



## METHODS

We describe here in detail the epitaxial growth and structural, magnetic and electronic characterization of the group III/nitride semiconductor heterostructures and  $\text{NbN}_x$  superconductors. We also describe the method of fabrication, as well as measurements and characterization, of the epitaxial semiconductor transistor/superconductor heterostructures and devices.

**MBE growth.** Epitaxial  $\text{NbN}_x$  films were grown at 800 °C by radio-frequency plasma-assisted MBE on three-inch-diameter, metal-polar semi-insulating 4H- and 6H-SiC substrates. The substrates had been commercially polished using chemical-mechanical polishing to an epi-ready finish, and were used as received. The reactive nitrogen was generated using a radio-frequency plasma source fed by ultrahigh-purity  $\text{N}_2$ , which was further purified by an in-line purifier. The Nb flux was generated using an *in situ* electron-beam evaporator source with 3N5-pure (excluding tantalum, Ta) Nb pellets in a tungsten hearth liner. Further details regarding MBE growth conditions are in ref. 25.

**Structural measurements.** We measured the surface morphology of the MBE-grown  $\text{NbN}_x$  films using a Bruker Dimension FastScan atomic force microscope in tapping mode. The root-mean-square roughness of the 5-nm  $\text{NbN}_x$  film on SiC is 0.15 nm in an area of  $3 \times 3 \mu\text{m}^2$ , and 0.56 nm for the 35-nm film (Extended Data Fig. 2). We determined the lattice constants and phase of the  $\text{NbN}_x$  films through X-ray diffraction (XRD) measurements, using a Rigaku system that employs a rotating copper anode to produce  $\text{Cu-K}\alpha$  radiation. Structural properties and lattice parameters of  $\text{NbN}_x$  on SiC are given in ref. 51. The measured XRD spectra of the 5-nm and 35-nm films are shown in Extended Data Fig. 3: the peaks for  $\text{NbN}_x$  are seen in first- and second-order reflection of the SiC (0004) plane in the relatively thick 35-nm sample, but the peaks are absent in the 5-nm sample because of the weak XRD signal in such thin films.

**Transport and magnetic measurements.** All of our electrical transport, Hall-effect and VSM measurements were carried out in a physical property measurement system (PPMS) manufactured by Quantum Design Inc. Extended Data Table 1 summarizes the basic material parameters extracted from measurements on the 5-nm and 35-nm samples. Extended Data Fig. 6 shows that, at 300 K, the carrier density in these  $\text{NbN}_x$  films is as high as  $n_{3d} \approx 10^{23} \text{ cm}^{-3}$ , but the mobility is less than  $1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ , probably limited by impurities, crystal defects and phonons. In terms of the superconducting behaviour, critical temperatures extracted from electrical resistance and VSM measurements are consistent with each other, but vary slightly from sample to sample in the range 6–15 K.

Characterization of the dependence of superconductivity on the in-plane magnetic field was limited by the 14 T field capabilities of our measurement system. We anticipate that our ongoing low-temperature and high-magnetic-field (up to 35 T) measurements will provide deeper insights into the orbital/spin pair-breaking mechanism and the presence of spin-orbit scattering.

**Critical current density.** Extended Data Fig. 6b, c shows the measured  $J_c$  in our MBE-grown  $\text{NbN}_x$  films. We carried out the measurements by injecting current into the film and detecting the voltage drop when the injected current density exceeded  $J_c$ . The measured values are close to  $10^5 \text{ A cm}^{-2}$  for MBE-grown  $\text{NbN}_x$ .

**Pair-breaking mechanisms in epitaxial thin  $\text{NbN}_x$  films.** The in-plane critical fields measured for the 5-nm  $\text{NbN}_x$  epitaxial superconducting layers are much higher than expected from the out-of-plane critical fields and the Tinkham formula. Mechanisms that could lead to this phenomenon include a modified electron  $g$ -factor<sup>31</sup>, the presence of spin-orbit scattering<sup>52</sup>, and Rashba spin-orbit coupling<sup>32</sup>.

If the mechanism were a modified electron  $g$ -factor, then given that the measured critical in-plane field is a factor of two larger than the Pauli limit, the epitaxial  $\text{NbN}_x$  would need an effective  $g$ -factor of less than 1, which we find unlikely: because  $\text{NbN}_x$  is a good metal, we suspect its effective  $g$ -factor to be close to 2 (ref. 53). Spin-orbit scattering is possible in the MBE-grown  $\text{NbN}_x$  films owing to the presence of trace amounts of Ta in the purest available Nb sources. However, the dilute concentration of Ta that we have measured in our MBE  $\text{NbN}_x$  suggests that this scenario is unlikely.

Finally, the presence of Rashba spin-splitting owing to broken inversion symmetry of the samples has recently been suggested<sup>32</sup> as a mechanism by which to suppress the Pauli paramagnetic limit. Because our films are grown in an asymmetric stack, we find this the most plausible explanation. Previous experimental and theoretical work has suggested the importance of Rashba spin-orbit coupling in identifying an anomalously large  $H_{c2}$ . However, our epitaxial  $\text{NbN}_x$  provides a platform for testing this idea directly, because of the ability to grow nominally symmetric stacks in  $\text{NbN}_x$ —a feat difficult in ultrathin lead films<sup>32</sup>, but potentially possible, if challenging, in two-dimensional materials.

**Shubnikov–de Haas oscillations.** Extended Data Fig. 7a shows the raw measured Shubnikov–de Haas oscillations of the GaN/AlGaIn 2DEG grown epitaxially on  $\text{NbN}_x$  layers. The oscillations become sharper as the temperature is lowered. We

used these Shubnikov–de Haas measurements, with a fit to the Lifshitz–Kosevich form of the magnetoresistance, to extract carrier concentration, effective mass and quantum-scattering times as discussed in the main text. The magnetoresistance data were uniformly resampled over an inverse magnetic field, and then smoothed over a window of  $0.00056 \text{ T}^{-1}$  before background subtraction. To extract the effective mass and quantum-scattering times, we removed the non-oscillating background component of the resistance and used the oscillatory components (Extended Data Fig. 7a inset). A non-oscillatory background of the form  $p(1/B) = a + b/B^{1/2} + c/B$  was subtracted from the  $R_{xx}$  data before fitting to the Lifshitz–Kosevich form.

Extended Data Fig. 7b shows a Landau plot of the Shubnikov–de Haas oscillation peaks. The range of magnetic fields used in this measurement, 0–14 T, allowed the Fermi level to fill 19–25 Landau levels at a fixed 2DEG density of  $n_{2d} \approx 1.2 \times 10^{13} \text{ cm}^{-2}$ .

**Superconductor/semiconductor transistor devices.** To fabricate the GaN HEMT structure (Extended Data Fig. 4a inset), we first grew 28-nm  $\text{NbN}_x$  on a 6H-SiC substrate by MBE; this was followed by nucleation with 22 nm AlN, two-step application of a  $1.3\text{-}\mu\text{m}$  GaN buffer layer, and then growth of 32-nm  $\text{Al}_{0.4}\text{Ga}_{0.6}\text{N}$  and 32-nm GaN channel at 700 °C. After the growth, ohmic contacts with Ti/Al/Ni/Au (20/100/10/50 nm) stacks were defined by optical lithography and electron-beam evaporation. Rapid thermal annealing at 850 °C produced ohmic contacts with a contact resistance of  $0.4 \Omega \text{ mm}^{-1}$ . Inductively coupled plasma etching with a  $\text{Cl}_2/\text{BCl}_3/\text{Ar}$  gas was then used to isolate separate HEMTs. To reduce the gate leakage current, we deposited a 10-nm-thick, high-K dielectric layer of  $\text{TiO}_2$  by atomic-layer deposition at 300 °C; this was followed by Pt/Au (30/200 nm) electron-beam evaporation to produce the gate metal stack. Finally, the  $\text{TiO}_2$  on top of the drain and source contacts were removed with fluorine-based plasma etching, and a second metalization of Ti/Pt/Au (25/25/400 nm) was performed.

Using fabricated van der Pauw structures, we performed Hall-effect measurements on the 2DEG at the GaN/ $\text{Al}_{0.4}\text{Ga}_{0.6}\text{N}$  interface. We determined the electron concentration to be  $1.3 \times 10^{13} \text{ cm}^{-2}$ , with a mobility of  $1,350 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  at room temperature and  $3,400 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  at 2 K, indicating that a high-quality 2DEG channel is achieved in these heterostructures and, more importantly, that processing did not lead to performance degradation.

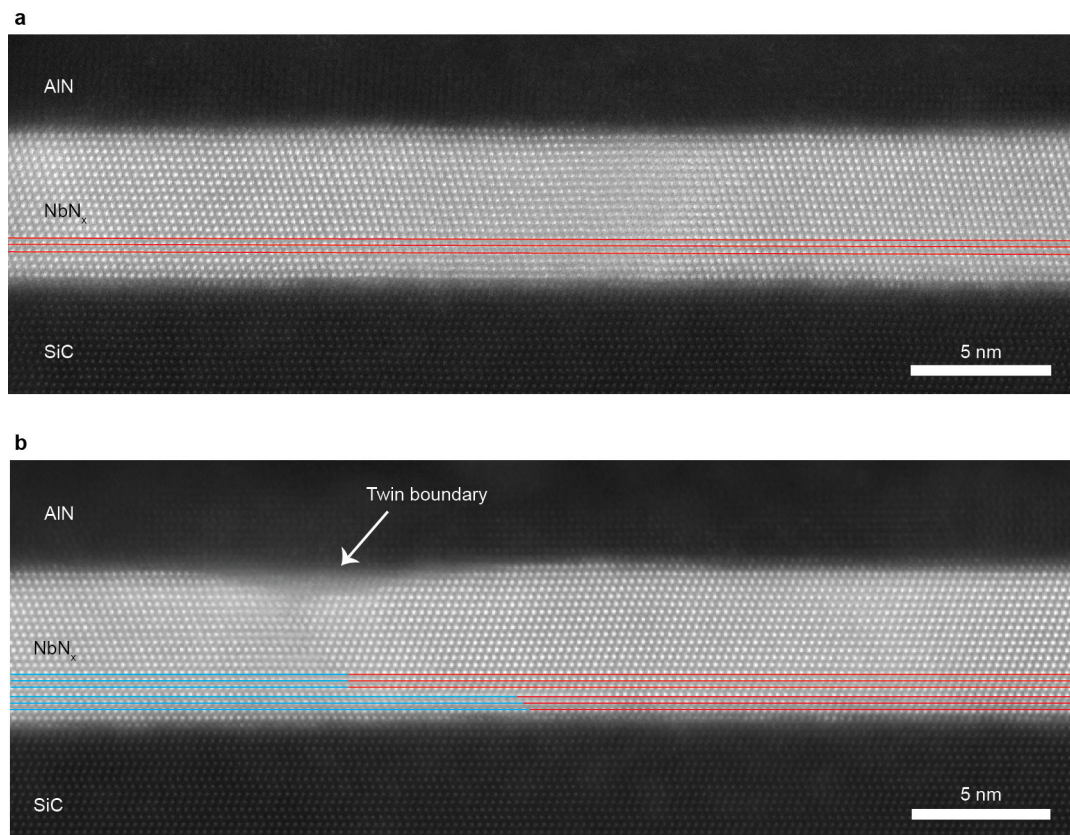
A representative room-temperature electrical characterization of the fabricated GaN HEMTs on  $\text{NbN}_x$  is shown in Extended Data Fig. 4. For a gate length of  $L_g = 1 \mu\text{m}$  and gate width of  $W = 75 \mu\text{m}$ , the transistors show an  $I_{\text{on}}/I_{\text{off}}$  ratio of more than  $10^5$  (Extended Data Fig. 4a). The on-current density exceeds  $1 \text{ A mm}^{-1}$  at  $V_d = 3 \text{ V}$  and  $V_g = -1 \text{ V}$ , with a clear current saturation (Extended Data Fig. 4b). Overall, the properties of the transistors studied here are similar to those of GaN HEMTs that are grown directly on SiC without the  $\text{NbN}_x$  layer underneath<sup>28</sup>. This is, to our knowledge, the first successful direct epitaxial integration of a high-performance semiconductor transistor on a superconductor.

Extended Data Fig. 5 shows the measured drain currents without the superconductor load ( $J_{ds}$ ; solid lines) and with the superconductor load ( $J_{ds}$ ; circles) at 5 K (less than  $T_c$ ) under two different drain voltages in a linear scale. We can see that, when  $V_{ds}$  and  $V_{ds'}$  are 0.1 V, the drain currents as a function of gate voltage are identical, because the  $\text{NbN}_x$  remains superconducting with a resistance of  $0 \Omega$  throughout this gate-voltage range. However, when  $V_{ds}$  and  $V_{ds'}$  are 1.1 V, the  $J_d$  versus  $V_{gs'}$  curve deviates from the  $J_d$  versus  $V_{gs}$  curve once  $J_d$  exceeds  $0.1 \text{ A mm}^{-1}$ . This indicates the occurrence of a superconductor-to-metal phase transition of the  $\text{NbN}_x$  film at the source end driven by this current (power) level.

**Determination of N/Nb ratio (x) by SIMS.** Extended Data Fig. 8 shows a SIMS measurement of the entire HEMT epitaxial heterostructure. Sharp and abrupt transitions of the SiC/superconductor, superconductor/AlN, AlN/GaN and AlGaIn/GaN heterointerfaces are observed. The SIMS profile provides a calibrated measurement of the stoichiometry of each layer of the heterostructure. The semiconducting AlN, GaN and AlGaIn layers are perfectly stoichiometric within the limits of the measurement, and the  $\text{NbN}_x$  layer has an N/Nb ratio (x) of  $43.3/56.7 = 0.762$ . Extended Data Table 2 shows additional N/Nb ratios measured by Rutherford back scattering (RBS) and SIMS, as well as the relation between N/Nb ratios, the residual resistance ratio (RRR) and the superconducting transition temperature.

**Data availability.** The datasets generated and analysed here are available from the corresponding author on reasonable request.

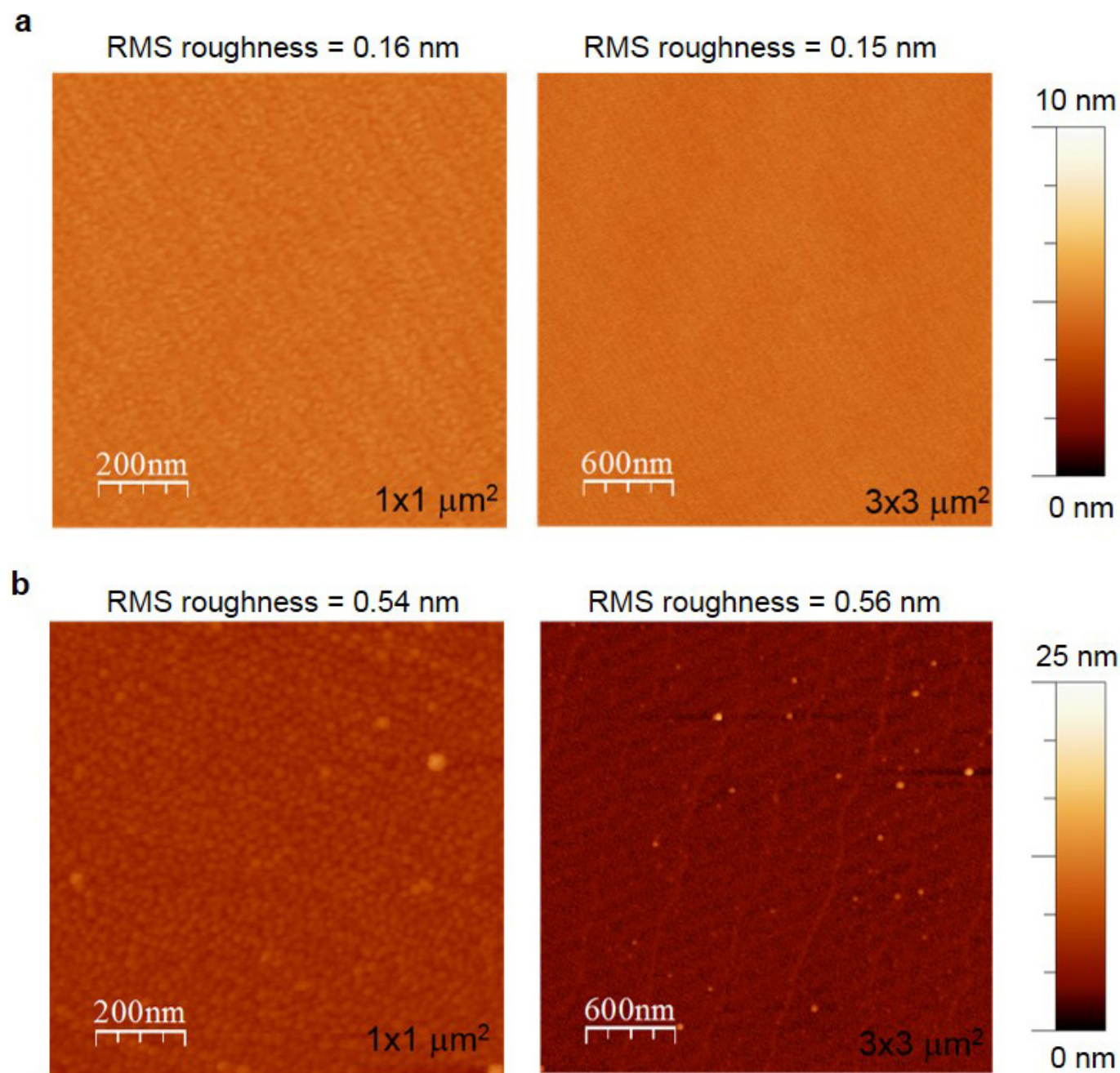
51. Nepal, N. *et al.* Characterization of molecular beam epitaxy grown  $\beta\text{-Nb}_2\text{N}$  films and AlN/ $\beta\text{-Nb}_2\text{N}$  heterojunctions on 6H-SiC substrates. *Appl. Phys. Exp.* **9**, 021003 (2016).
52. Werthamer, N., Helfand, E. & Hohenberg, P. Temperature and purity dependence of the superconducting critical field  $H_{c2}$ . III. Electron spin and spin-orbit effects. *Phys. Rev.* **147**, 295 (1966).
53. MacDonald, A. Transition-metal  $g$  factor trends. *J. Phys. F* **12**, 2579 (1982).



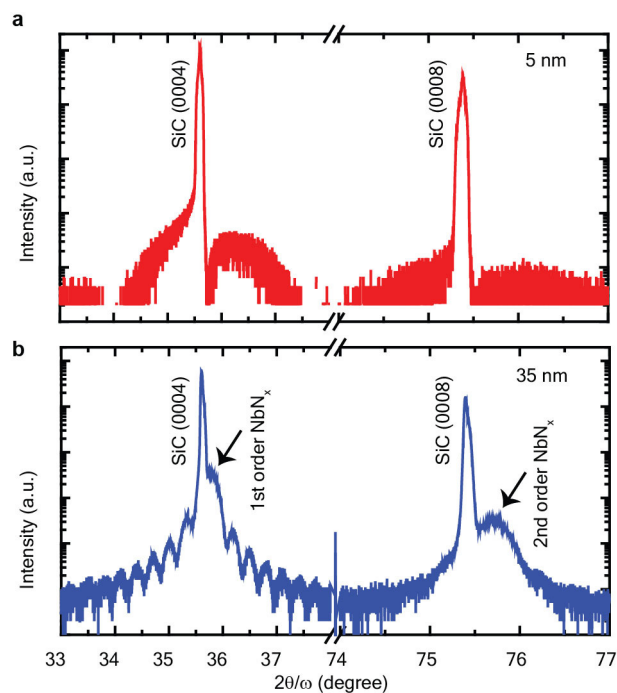
**Extended Data Figure 1 | STEM of a large-area, MBE-grown AlN/NbN<sub>x</sub>/SiC heterostructure.** **a**, STEM image of NbN<sub>x</sub>/AlN grown on top of a SiC substrate, showing the single-crystal nature of NbN<sub>x</sub> over a large region. The red lines have been added as a guide to show the crystallinity across the entire range measured. **b**, A twin boundary and a stacking fault in the

MBE NbN<sub>x</sub> layer. This STEM image of NbN<sub>x</sub>/AlN on top of SiC shows a grain boundary with two cubic NbN<sub>x</sub> phases rotated across each other. The red and blue lines have been added to draw out the stacking fault seen near the twin boundary.



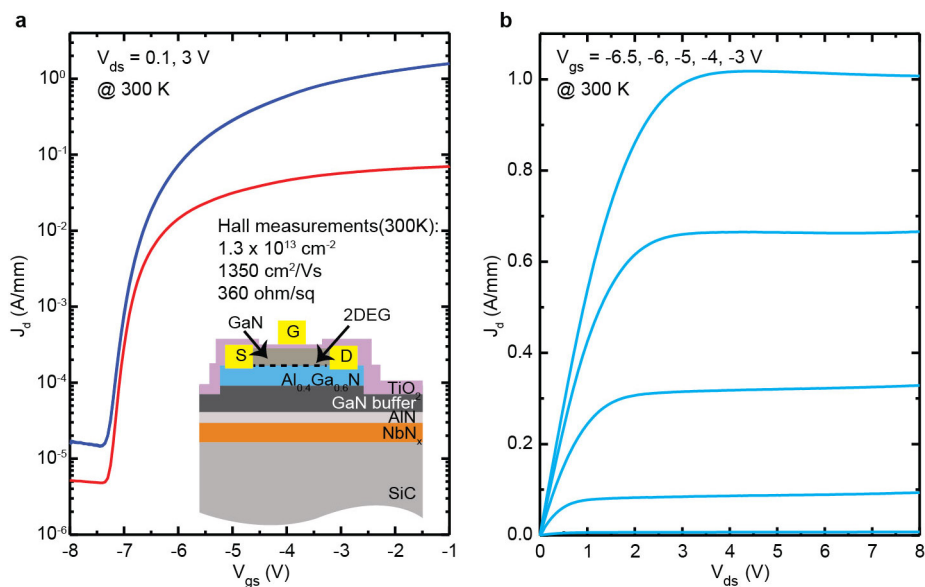


**Extended Data Figure 2 | AFM characterizations of thin films.** **a, b,** AFM images of epitaxial  $\text{NbN}_x$  films that are 5-nm thick (**a**) and 35-nm thick (**b**), over areas of  $1 \times 1 \mu\text{m}^2$  and  $3 \times 3 \mu\text{m}^2$ . RMS, root mean squared.



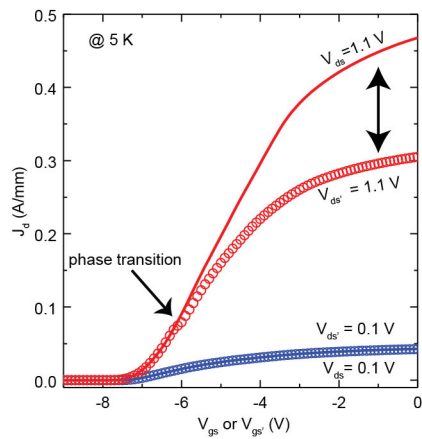
**Extended Data Figure 3 | Symmetrical  $2\theta/\omega$  XRD curves of 5-nm and 35-nm NbN<sub>x</sub> on 4H-SiC.**  $2\theta$  is the angle between the incident and diffracted beams, and  $\omega$  is the angle between the incident beam and the sample surface. **a**, The 5-nm sample. **b**, The 35-nm sample. There is a clear separation between the SiC and cubic NbN (first- and second-order) peaks in the 35-nm sample. But this feature is absent in the 5-nm sample, owing to the weak XRD signal intensity in such an ultrathin film.





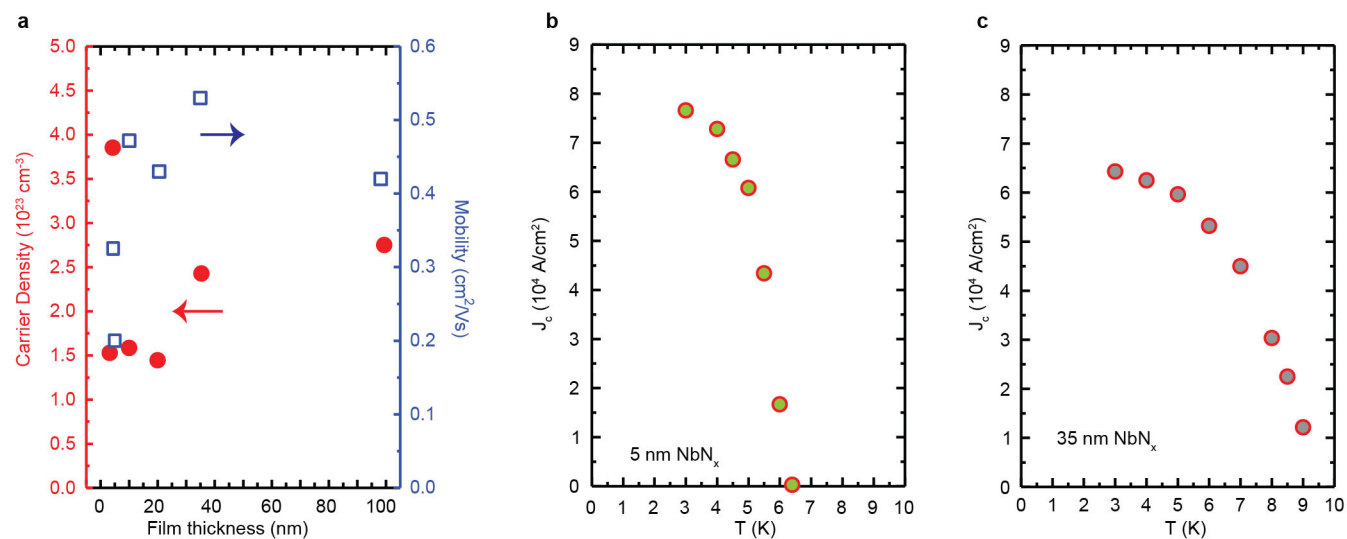
**Extended Data Figure 4 | Electrical characterizations of HEMT structures.** **a**, Drain current ( $J_d$ ) versus top-gate voltage ( $V_{gs}$ ) transfer curves at 300 K for HEMTs grown on NbN<sub>x</sub>/SiC substrate, showing a high on/off ratio at source–drain voltages ( $V_{ds}$ ) of 0.1 V (red curve) and 3 V

(blue curve). **b**,  $J_d$  versus  $V_{ds}$  curves for various top-gate voltages (from bottom curve to top,  $-3 \text{ V}$ ,  $-4 \text{ V}$ ,  $-5 \text{ V}$ ,  $-6 \text{ V}$ ,  $-6.5 \text{ V}$ ) of GaN HEMTs at 300 K.



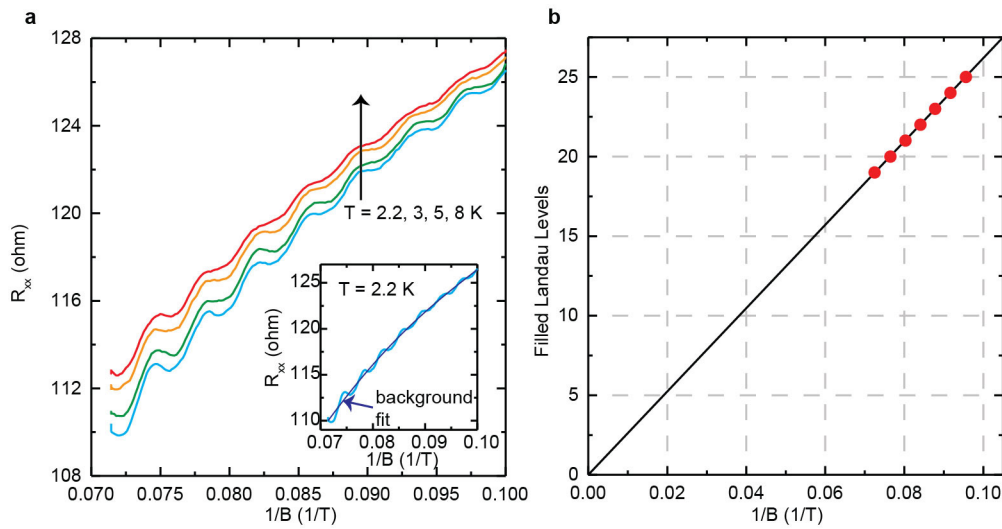
**Extended Data Figure 5 | Representative transfer curves of a HEMT structure with superconducting load.** The graph plots  $J_d$  versus  $V_{gs}$  (without superconductor load) and  $V_{gs'}$  (with superconductor load) at 5 K, showing the phase transition of  $NbN_x$  that occurs when  $J_d$  is larger than  $0.1 \text{ A mm}^{-1}$ .





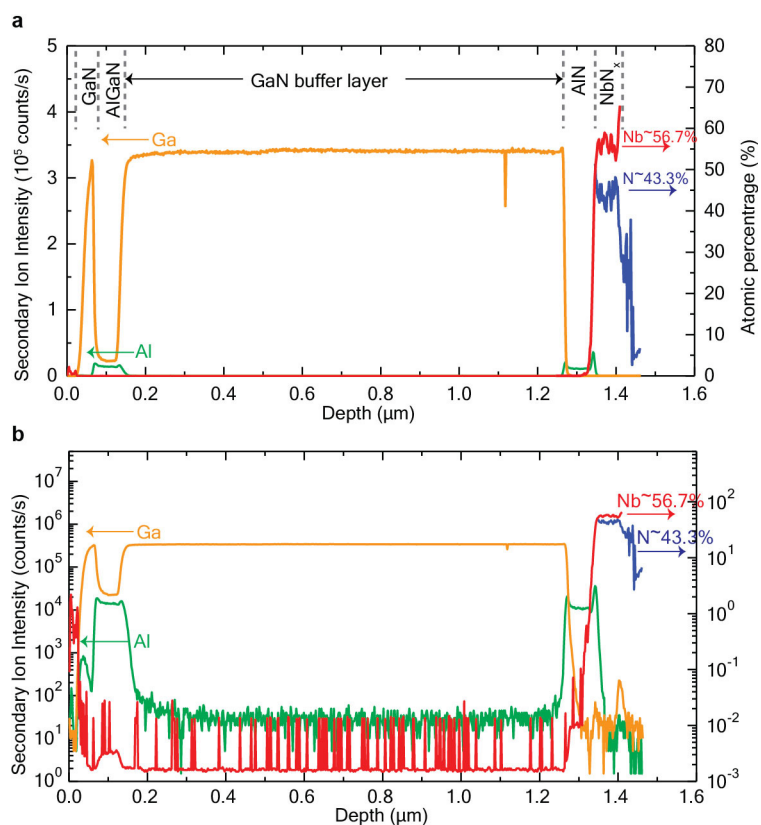
**Extended Data Figure 6 | Electrical characterizations of superconducting  $\text{NbN}_x$  films.** **a**, Summary of carrier density and mobility for different thicknesses of  $\text{NbN}_x$  films, ranging from 4 nm to 100 nm. The red arrow

shows that the red dots correspond to the left-hand y axis; the blue squares correspond to the right-hand y axis. **b**, **c**, Critical current density as a function of temperature for 5-nm and 35-nm  $\text{NbN}_x$  films.



**Extended Data Figure 7 | Shubnikov-de Haas oscillations of 2DEG.**  
**a**, Magnetoresistance ( $R_{xx}$ ) plotted against inverse magnetic field ( $1/B$ ) before background subtraction, taken over the magnetic-field range of 10 T to 14 T. The oscillations occur at periods of  $1/B$ —a clear indication of sharp peaks in the 2DEG density of states owing to Landau levels. The upward black arrow indicates increasing temperatures. The inset shows

the 2.2 K data (light blue) with background (blue); the non-oscillating background was removed before evaluation of carrier concentration, effective mass, and scattering time. **b**, Landau plot of magnetoresistance relative minima plotted against inverse magnetic field. The slope of the line is  $\hbar^2 \pi n_{\text{SDH}} / 2m^*$ ; the density and effective mass are taken from the Lifshitz-Kosevich fit to magnetoresistance oscillations.



**Extended Data Figure 8 | SIMS results obtained on the SiC/NbN<sub>x</sub>/AlN/GaN/AlGaN/GaN heterostructure.** The left-hand y axis shows the secondary-ion intensity (counts per second) of Al and Ga atoms. The right-hand y axis indicates the atomic percentages of N and Nb atoms,

which are respectively 43.3% and 56.7%, indicating a N/Nb ratio ( $x$ ) of 0.762 in linear (a) and log (b) scale of signal intensity. The  $x$  axis denotes the depth of the sample from the top surface.



Extended Data Table 1 | Characteristic quantities in the normal state and superconducting state for thick and thin epitaxial NbN<sub>x</sub> films on 4H-SiC at 15 K

Qauntity	Symbol	Units	Thin Film	Thick Film
Film thicknness	$d$	nm	5	35
Sheet resistance	$R_{sh}$	$\Omega$	107.9	9.8
Carrier concentration	$n_{3d}$	$10^{23} \text{ cm}^{-3}$	2.14	2.29
Hall effect mobility	$\mu$	$\text{cm}^{-2}/\text{Vs}$	0.54	0.93
Mean free path	$l_{MFP}$	$\text{\AA}$	6.6	11.6
Tc (Resistance)	$T_c$	K	6.41	9.26
Tc (Magnetometry)	$T_c$	K	6.23	9.36
Coherence length	$\xi$	nm	10.56	10.06

Extended Data Table 2 | Epitaxial NbN<sub>x</sub> film properties

Sample	Thickness (nm)	<i>RRR</i>	<i>T<sub>c</sub></i> (K)	N/Nb ratio ( <i>x</i> )	Notes
A	5	1.50	6.41	0.877 (RBS)	NbN <sub>x</sub> on SiC
B	35	1.86	9.26	0.751 (RBS)	NbN <sub>x</sub> on SiC
C	28	1.69	7.70	0.762 (SIMS)	GaN/AlGaN/GaN/AlN/NbN <sub>x</sub> on SiC

Film thickness, residual resistance ratio (*RRR*), superconducting critical temperature (*T<sub>c</sub>*) and N/Nb ratio (*x*) are given for the samples discussed here. The film thickness was determined by RBS and TEM. *RRR* is defined as the ratios of the normal-state resistance at 300 K and at 20 K. *T<sub>c</sub>* is identified as the midpoint of the superconducting transition. The N/Nb ratio *x* was measured by either RBS or SIMS.

# The Beaker phenomenon and the genomic transformation of northwest Europe

A list of authors and affiliations appears at the end of the paper.

**From around 2750 to 2500 BC, Bell Beaker pottery became widespread across western and central Europe, before it disappeared between 2200 and 1800 BC. The forces that propelled its expansion are a matter of long-standing debate, and there is support for both cultural diffusion and migration having a role in this process. Here we present genome-wide data from 400 Neolithic, Copper Age and Bronze Age Europeans, including 226 individuals associated with Beaker-complex artefacts. We detected limited genetic affinity between Beaker-complex-associated individuals from Iberia and central Europe, and thus exclude migration as an important mechanism of spread between these two regions. However, migration had a key role in the further dissemination of the Beaker complex. We document this phenomenon most clearly in Britain, where the spread of the Beaker complex introduced high levels of steppe-related ancestry and was associated with the replacement of approximately 90% of Britain's gene pool within a few hundred years, continuing the east-to-west expansion that had brought steppe-related ancestry into central and northern Europe over the previous centuries.**

During the third millennium BC, two new archaeological pottery styles expanded across Europe and replaced many of the more localized styles that had preceded them<sup>1</sup>. The expansion of the 'Cord Ware complex' in north-central and northeastern Europe was associated with people who derived most of their ancestry from populations related to Early Bronze Age Yamnaya pastoralists from the Eurasian steppe<sup>2–4</sup> (henceforth referred to as 'steppe'). In western Europe there was the equally expansive 'Bell Beaker complex', defined by assemblages of grave goods that included stylized bell-shaped pots, copper daggers, arrowheads, stone wristguards and V-perforated buttons<sup>5</sup> (Extended Data Fig. 1). The oldest radiocarbon dates associated with Beaker pottery are from around 2750 BC in Atlantic Iberia<sup>6</sup>, which has been interpreted as evidence that the Beaker complex originated in this region. However, the geographic origins of this complex are still debated<sup>7</sup> and other scenarios—including an origin in the Lower Rhine area, or even multiple independent origins—are possible (Supplementary Information section 1). Regardless of geographic origin, by 2500 BC the Beaker complex had spread throughout western Europe and northwest Africa and had reached southern and Atlantic France, Italy and central Europe<sup>5</sup>, where it overlapped geographically with the Cord Ware complex. Within another hundred years, it had expanded to Britain and Ireland<sup>8</sup>. A major debate in archaeology has revolved around the question of whether the spread of the Beaker complex was mediated by the movement of people, culture or a combination of both<sup>9</sup>. Genome-wide data have revealed high proportions of steppe-related ancestry in Beaker-complex-associated individuals from Germany and the Czech Republic<sup>2–4</sup>, which shows that these individuals derived from mixtures of populations from the steppe and the preceding Neolithic farmers of Europe. However, a deeper understanding of the ancestry of people associated with the Beaker complex requires genomic characterization of individuals across the geographic range and temporal duration of this archaeological phenomenon.

## Ancient DNA data

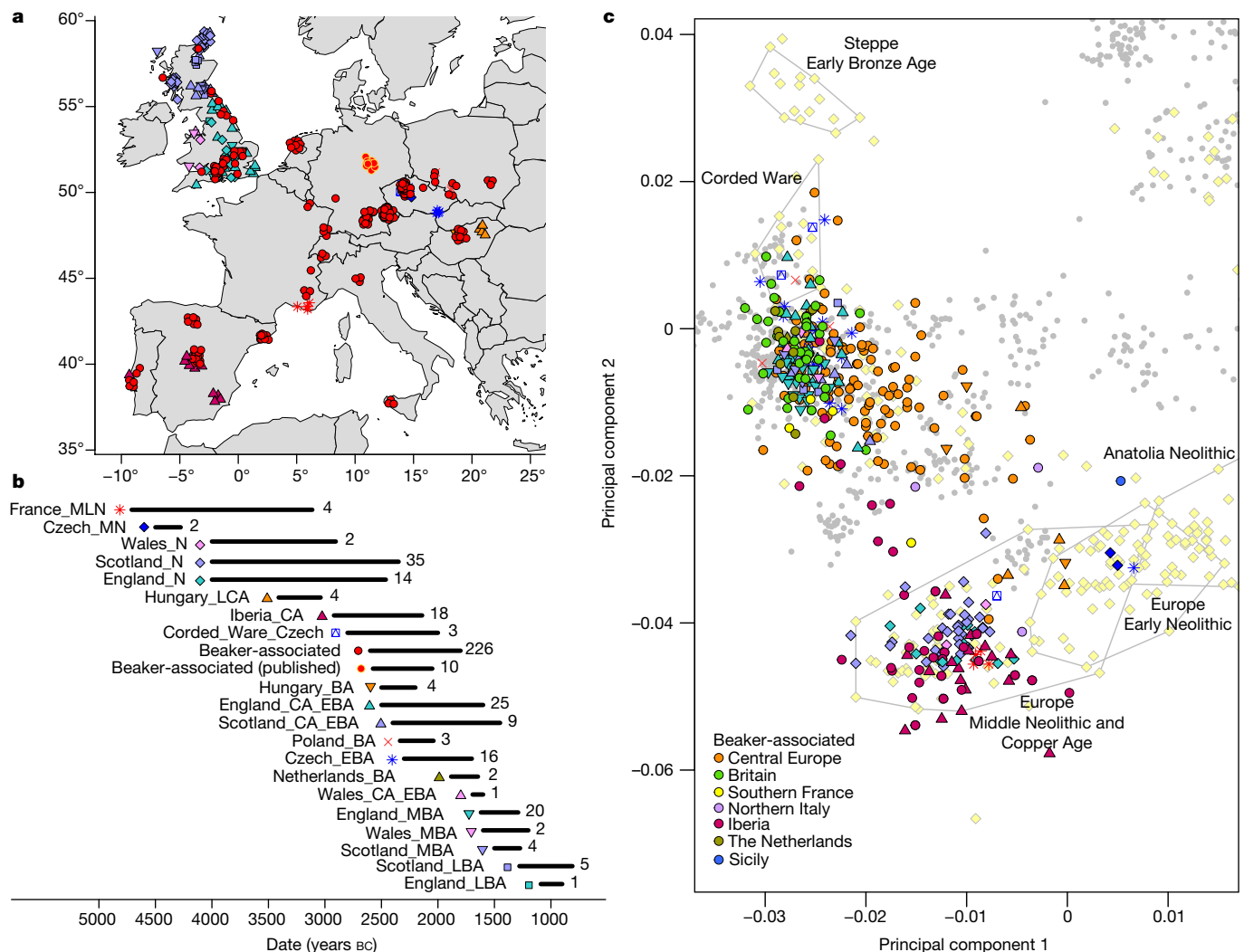
To understand the genetic structure of ancient people associated with the Beaker complex and their relationship to preceding, subsequent and contemporary peoples, we used hybridization DNA capture<sup>4,10</sup> to enrich

ancient DNA libraries for sequences overlapping 1,233,013 single nucleotide polymorphisms (SNPs), and generated new sequence data from 400 ancient Europeans dated to between approximately 4700 and 800 BC, excavated from 136 different sites (Extended Data Tables 1, 2; Supplementary Table 1; Supplementary Information section 2). This dataset includes 226 Beaker-complex-associated individuals from Iberia ( $n = 37$ ), southern France ( $n = 4$ ), northern Italy ( $n = 3$ ), Sicily ( $n = 3$ ), central Europe ( $n = 133$ ), the Netherlands ( $n = 9$ ) and Britain ( $n = 37$ ), and 174 individuals from other ancient populations, including 118 individuals from Britain who lived both before ( $n = 51$ ) and after ( $n = 67$ ) the arrival of the Beaker complex (Fig. 1a, b). For genome-wide analyses, we filtered out first-degree relatives and individuals with low coverage (fewer than 10,000 SNPs) or evidence of DNA contamination (Methods) and combined our data with previously published ancient DNA data (Extended Data Fig. 2) to form a dataset of 683 ancient samples (Supplementary Table 1). We merged these data with those from 2,572 present-day individuals genotyped on the Affymetrix Human Origins array<sup>11,12</sup> as well as with 300 high-coverage genomes<sup>13</sup>. To facilitate the interpretation of our genetic results, we also generated 111 direct radiocarbon dates (Extended Data Table 3; Supplementary Information section 3).

## Y-chromosome analysis

The Y-chromosome composition of Beaker-complex-associated males was dominated by R1b-M269 (Supplementary Table 4), which is a lineage associated with the arrival of steppe migrants in central Europe after 3000 BC<sup>2,3</sup>. Outside Iberia, this lineage was present in 84 out of 90 analysed males. For individuals for whom we determined the R1b-M269 subtype ( $n = 60$ ), we found that all but two had the derived allele for the R1b-S116/P312 polymorphism, which defines the dominant subtype in western Europe today<sup>14</sup>. By contrast, Beaker-complex-associated individuals from the Iberian Peninsula carried a higher proportion of Y haplogroups known to be common across Europe during the earlier Neolithic period<sup>2,4,15,16</sup>, such as I ( $n = 5$ ) and G2 ( $n = 1$ ); R1b-M269 was found in four individuals with a genome-wide signal of steppe-related ancestry, and of these, the two with higher coverage could be classified





**Figure 1 | Spatial, temporal and genetic structure of individuals in this study.** **a**, Geographic distribution of samples with new genome-wide data. Random jitter was added for sites with multiple individuals. Map data from the R package 'maps'. **b**, Approximate time ranges for samples with new genome-wide data. Sample sizes are given next to each bar.

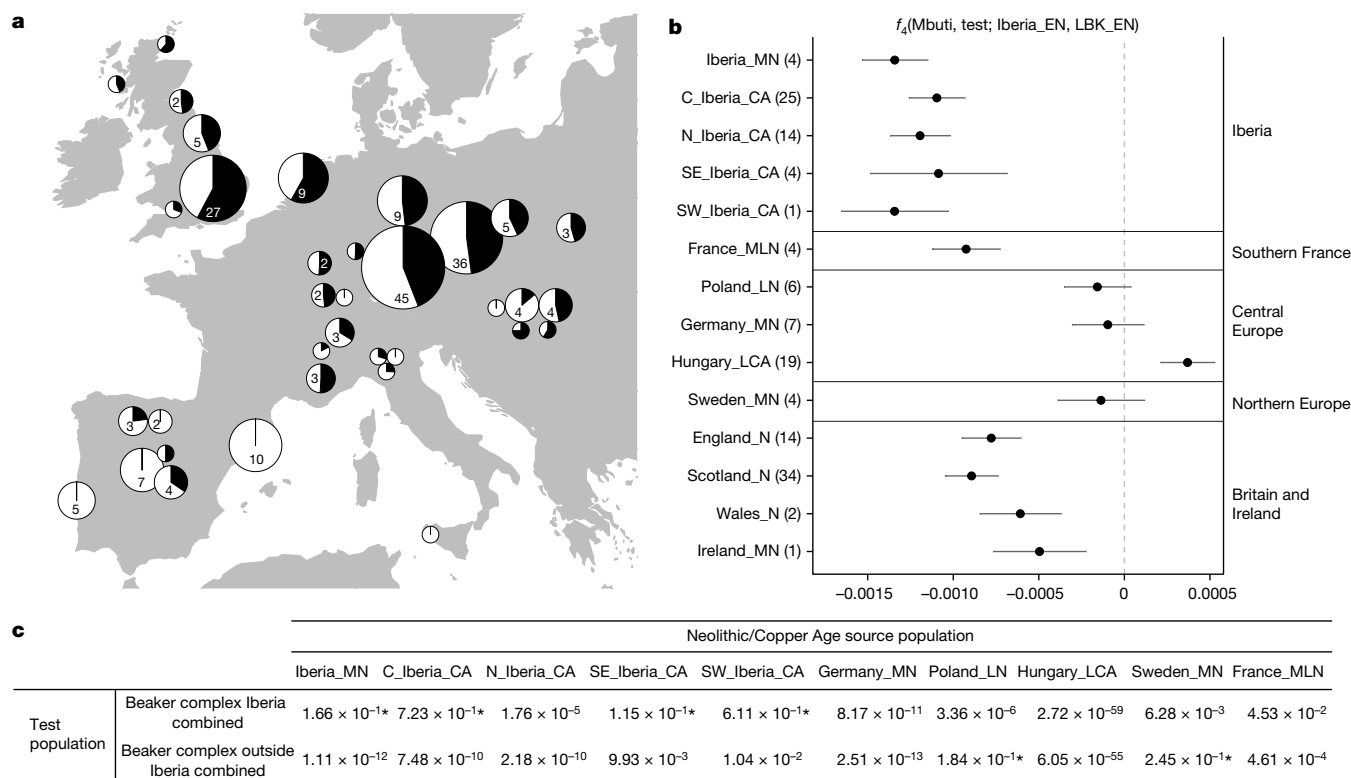
as R1b-S116/P312. The widespread presence of the R1b-S116/P312 polymorphism in ancient individuals from central and western Europe suggests that people associated with the Beaker complex may have had an important role in the dissemination of this lineage throughout most of its present-day distribution.

### Spread of people associated with the Beaker complex

We performed principal component analysis by projecting the ancient samples onto the genetic variation in a set of west Eurasian present-day populations. We replicated previous findings<sup>11</sup> of two parallel clines, with present-day Europeans on one side and present-day Near Eastern populations on the other (Extended Data Fig. 3a). Individuals associated with the Beaker complex are notably heterogeneous within the European cline along an axis of variation defined by Early Bronze Age Yamnaya individuals from the steppe at one extreme and Middle Neolithic and Copper Age Europeans at the other extreme (Fig. 1c; Extended Data Fig. 3a). This suggests that genetic differentiation among Beaker-complex-associated individuals may be related to variable amounts of steppe-related ancestry. We obtained qualitatively consistent inferences using ADMIXTURE model-based clustering<sup>17</sup>. Beaker-complex-associated individuals harboured three main genetic components: one characteristic of European Mesolithic

hunter-gatherers, one maximized in Neolithic individuals from the Levant and Anatolia, and one maximized in Neolithic individuals from Iran and present in admixed form in steppe populations (Extended Data Fig. 3b).

Both principal component analysis and ADMIXTURE are powerful tools for visualizing genetic structure, but they do not provide formal tests of admixture between populations. We grouped Beaker-complex-associated individuals on the basis of geographic proximity and genetic similarity (Supplementary Information section 6), and used qpAdm<sup>2</sup> to directly test admixture models and estimate mixture proportions. We modelled their ancestry as a mixture of Mesolithic western European hunter-gatherers, northwestern Anatolian Neolithic farmers and Early Bronze Age steppe populations; the first two of these contributed to the ancestry of earlier Neolithic Europeans. We find that in areas outside of Iberia, with the exception of Sicily, a large majority of the Beaker-complex-associated individuals that we sampled derive a considerable portion of their ancestry from steppe populations (Fig. 2a). By contrast, in Iberia such ancestry is present in only 8 of the 32 individuals that we analysed; these individuals represent the earliest detection of steppe-related genomic affinities in this region. We observed differences in ancestry not only at a pan-European scale, but also within regions and even within sites. For instance, at Szigetszentmiklós in Hungary, we



**Figure 2 | Investigating the genetic makeup of Beaker-complex-associated individuals.** **a**, Proportion of steppe-related ancestry (in black) in Beaker-complex-associated groups computed with qpAdm<sup>2</sup> under the model ‘Steppe\_EBA + Anatolia\_N + WHG’ (WHG, Mesolithic western European hunter-gatherers). The area of the pie is proportional to the number of individuals (number shown if more than one). Map data from the R package ‘maps’. **b**,  $f_4$ -statistics of the form  $f_4(\text{Mbuti, test; Iberia\_EN, LBK\_EN})$

LBK\_EN) computed for European populations (number of individuals for each group is given in parentheses) before the emergence of the Beaker complex (Supplementary Information section 7). Error bars represent  $\pm 1$  standard errors. **c**, Testing different populations as a source for the Neolithic ancestry component in Beaker-complex-associated individuals. The table shows  $P$  values (\* indicates values  $> 0.05$ ) for the fit of the model: ‘Steppe\_EBA + Neolithic/Copper Age’ source population.

found roughly contemporary Beaker-complex-associated individuals with very different proportions (from 0% to 75%) of steppe-related ancestry. This genetic heterogeneity is consistent with early stages of mixture between previously established European Neolithic populations and migrants with steppe-related ancestry. One implication of this is that even at local scales, the Beaker complex was associated with people of diverse ancestries.

Although the steppe-related ancestry in Beaker-complex-associated individuals had a recent origin in the east<sup>2,3</sup>, the other ancestry component—from previously established European populations—could potentially be derived from several parts of Europe, because groups that were genetically closely related were widely distributed during the Neolithic and Copper Ages<sup>2,4,11,16,18–23</sup>. To obtain insight into the origin of this ancestry component in Beaker-complex-associated individuals, we looked for regional patterns of genetic differentiation within Europe during the Neolithic and Copper Age. We examined whether populations pre-dating the emergence of the Beaker complex shared more alleles with Iberian (Iberia\_EN) or central European Linearbandkeramik (LBK\_EN) Early Neolithic populations (Fig. 2b). As previously described<sup>2</sup>, Iberian Middle Neolithic and Copper Age populations, but not central and northern European populations, had genetic affinities with Iberian Early Neolithic farmers (Fig. 2b). These regional patterns could partially be explained by differential genetic affinities to pre-Neolithic hunter-gatherer individuals from different regions<sup>22</sup> (Extended Data Fig. 4). Neolithic individuals from southern France and Britain are also significantly closer to Iberian Early Neolithic farmers than they are to central European Early Neolithic farmers (Fig. 2b), consistent with a previous analysis of a Neolithic genome from Ireland<sup>23</sup>. By modelling Neolithic populations and Mesolithic western European

hunter-gatherers in an admixture graph framework, we replicate these results and show that they are not driven by different proportions of hunter-gatherer admixture (Extended Data Fig. 5; Supplementary Information section 7). Our results suggest that a portion of the ancestry of the Neolithic populations of Britain was derived from migrants who spread along the Atlantic coast. Megalithic tombs document substantial interaction along the Atlantic façade of Europe<sup>24</sup>, and our results are consistent with such interactions reflecting south-to-north movements of people. More data from southern Britain and Ireland and nearby regions in continental Europe will be needed to fully understand the complex interactions between Britain, Ireland and the continent during the Neolithic<sup>24</sup>.

The distinctive genetic signatures found in the Iberian populations who preceded the arrival of Beaker complex, when compared to contemporary central European populations, enable us to formally test for the origin of the Neolithic-related ancestry in Beaker-complex-associated individuals. We grouped individuals from Iberia ( $n = 32$ ) and from outside Iberia ( $n = 172$ ) to increase power and evaluated the fit of different Neolithic and Copper Age groups with qpAdm<sup>2</sup> under the model: ‘Steppe\_EBA + Neolithic/Copper Age’. For Beaker-complex-associated individuals from Iberia, the best fit was obtained when Middle Neolithic and Copper Age populations from the same region were used as the source for their Neolithic-related ancestry; we could exclude central and northern European populations as sources of this ancestry ( $P < 0.0063$ ) (Fig. 2c). Conversely, the Neolithic-related ancestry in Beaker-complex-associated individuals outside of Iberia was most closely related to central and northern European Neolithic populations with relatively high hunter-gatherer admixture (for example, Poland\_LN,  $P = 0.18$  and Sweden\_MN,  $P = 0.25$ ), and we could significantly exclude Iberian sources ( $P < 0.0104$ ) (Fig. 2c).

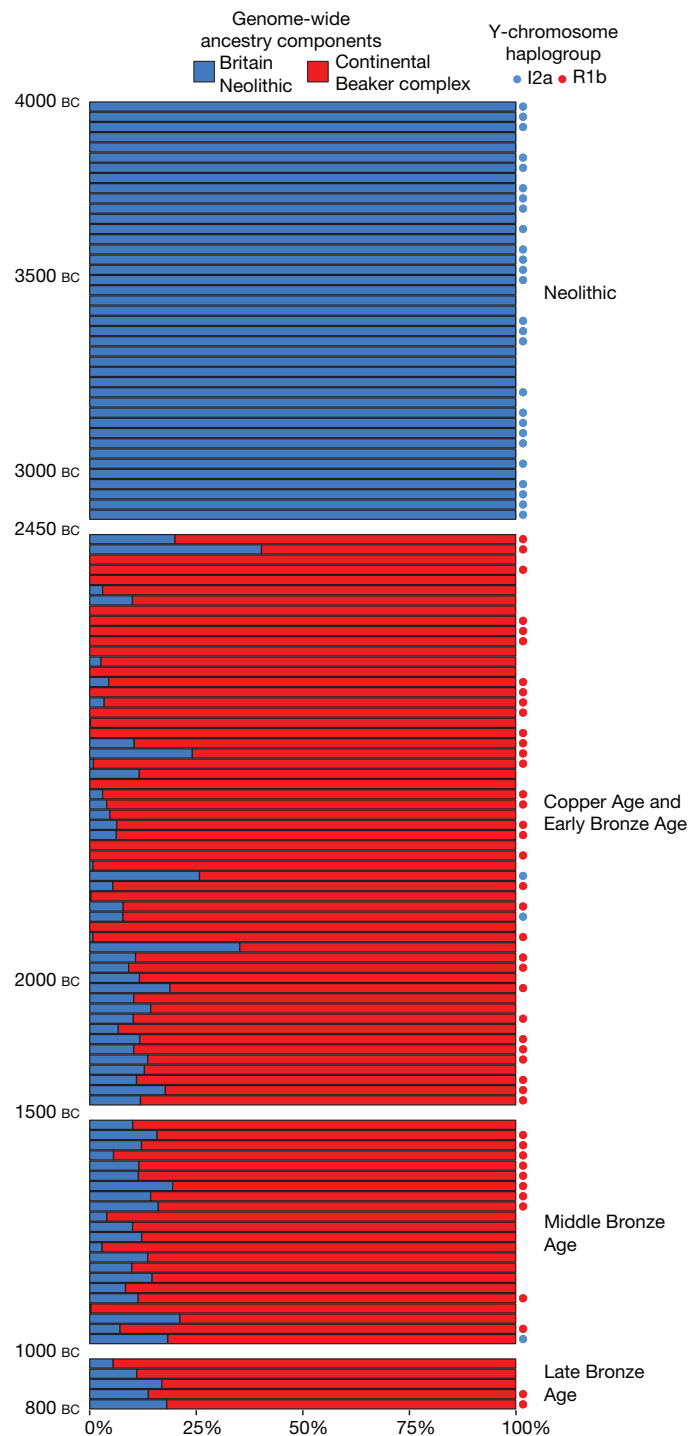
These results support mostly different origins for Beaker-complex-associated individuals, with no discernible Iberia-related ancestry outside of Iberia.

### Nearly complete turnover of ancestry in Britain

The genetic profile of British Beaker-complex-associated individuals ( $n = 37$ ) shows strong similarities to that of central European Beaker-complex-associated individuals (Extended Data Fig. 3). This observation is not restricted to British individuals associated with the 'All-Over-Cord' Beaker pottery style that is shared between Britain and central Europe: we also find this genetic signal in British individuals associated with Beaker pottery styles derived from the 'Maritime' forms, which were predominant earlier in Iberia. The presence of large amounts of steppe-related ancestry in British Beaker-complex-associated individuals (Fig. 2a) contrasts sharply with Neolithic individuals from Britain ( $n = 51$ ), who have no evidence of steppe genetic affinities and cluster instead with Middle Neolithic and Copper Age populations from mainland Europe (Extended Data Fig. 3). A previous study showed that steppe-related ancestry had arrived in Ireland by the Bronze Age<sup>23</sup>; here we show that, at least in Britain, it arrived earlier in the Copper Age (which, in Britain, is synonymous with the Beaker period).

Among the continental Beaker-complex groups analysed in our dataset, individuals from Oostwoud, the Netherlands, are the most closely related to the large majority of Beaker-complex-associated individuals from southern Britain ( $n = 27$ ). The two groups had almost identical steppe-related ancestry proportions (Fig. 2a), the highest level of shared genetic drift (Extended Data Fig. 6b) and were symmetrically related to most ancient populations (Extended Data Fig. 6a), which shows that they are likely derived from the same ancestral population with limited mixture into either group. This does not necessarily imply that the Oostwoud individuals are direct ancestors of the British individuals, but it does show that they were closely related genetically to the population—perhaps yet to be sampled—that moved into Britain from continental Europe.

We investigated the magnitude of population replacement in Britain with qpAdm<sup>2</sup> by modelling the genome-wide ancestry of Neolithic, Copper and Bronze Age individuals, including Beaker-complex-associated individuals, as a mixture of continental Beaker-complex-associated samples (using the Oostwoud individuals as a surrogate) and the British Neolithic population (Supplementary Information section 8). During the first centuries after the initial contact, between approximately 2450 and 2000 BC, ancestry proportions were variable (Fig. 3), which is consistent with migrant communities just beginning to mix with the previously established British Neolithic population. After roughly 2000 BC, individuals were more homogeneous and possessed less variation in ancestry proportions and a modest increase in Neolithic-related ancestry (Fig. 3). This could represent admixture with persisting British populations with high levels of Neolithic-related ancestry or, alternatively, with incoming continental populations with higher proportions of Neolithic-related ancestry. In either case, our results imply a minimum of  $90 \pm 2\%$  local population turnover by the Middle Bronze Age (approximately 1500–1000 BC), with no significant decrease observed in 5 samples from the Late Bronze Age. Although the exact turnover rate and its geographic pattern await refinement with more ancient samples, our results imply that for individuals from Britain during and after the Beaker period, a very high fraction of their DNA derives from ancestors who lived in continental Europe before 2450 BC. An independent line of evidence for population turnover comes from uniparental markers. Whereas Y-chromosome haplogroup R1b was completely absent in Neolithic individuals ( $n = 33$ ), it represents more than 90% of the Y chromosomes in individuals from Copper and Bronze Age Britain ( $n = 52$ ) (Fig. 3). The introduction of new mtDNA haplogroups, such as I, R1a and U4, which were present in Beaker-complex-associated populations from continental Europe but not in



**Figure 3 | Population transformation in Britain associated with the arrival of the Beaker complex.** Modelling Neolithic, Copper and Bronze Age (including Beaker-complex-associated) individuals from Britain as a mixture of continental Beaker-complex-associated individuals (red) and the Neolithic population from Britain (blue). Each bar represents genome-wide mixture proportions for one individual. Individuals are ordered chronologically and included in the plot if represented by more than 100,000 SNPs. Circles indicate the Y-chromosome haplogroup for male individuals.

Neolithic Britain (Supplementary Table 3), suggests that both men and women were involved in this population turnover.

Our ancient DNA transect-through-time in Britain also enabled us to track the frequencies of alleles with known phenotypic effects. Derived alleles at rs16891982 in *SLC45A2* and rs12913832 in *HERC2/OCA2*,



which contribute to reduced skin and eye pigmentation in Europeans, considerably increased in frequency between the Neolithic period and the Beaker and Bronze Age periods (Extended Data Fig. 7). The arrival of migrants associated with the Beaker complex therefore markedly altered the pigmentation phenotypes of British populations. However, the lactase persistence allele at SNP rs4988235 in *LCT* remained at very low frequencies across this transition, both in Britain and continental Europe, which shows that the major increase in its frequency occurred within the last 3,500 years<sup>3,4,25</sup>.

## Discussion

The term 'Bell Beaker' was introduced by late-nineteenth- and early-twentieth-century archaeologists to refer to a distinctive pottery style found across western and central Europe at the end of the Neolithic that was initially hypothesized to have been spread by a genetically homogeneous population. This idea of a 'Beaker Folk' became unpopular after the 1960s as scepticism grew about the role of migration in mediating change in archaeological cultures<sup>26</sup>, although even at the time<sup>27</sup> it was speculated that the expansion of the Beaker complex into Britain was an exception—a prediction that has now been borne out by ancient genomic data.

The expansion of the Beaker complex cannot be described by a simple one-to-one mapping of an archaeologically defined material culture to a genetically homogeneous population. This stands in contrast to other archaeological complexes, notably the Linearbandkeramik farmers of central Europe<sup>2</sup>, the Early Bronze Age Yamnaya of the steppe<sup>2,3</sup> and—to some extent—the Corded Ware complex of central and eastern Europe<sup>2,3</sup>. Our results support a model in which cultural transmission and human migration both had important roles, with the relative balance of these two processes depending on the region. In Iberia, the majority of Beaker-complex-associated individuals lacked steppe affinities and were genetically most similar to preceding Iberian populations. In central Europe, steppe-related ancestry was widespread and we can exclude a substantial contribution from Iberian Beaker-complex-associated individuals. However, the presence of steppe-related ancestry in some Iberian individuals demonstrates that gene flow into Iberia was not uncommon during this period. These results contradict initial suggestions of gene flow into central Europe based on analysis of mtDNA<sup>28</sup> and dental morphology<sup>29</sup>. In particular, mtDNA haplogroups H1 and H3 were proposed as markers for a Beaker-complex expansion originating in Iberia<sup>28</sup>, yet H3 is absent among our Iberian Beaker-complex-associated individuals.

In other parts of Europe, the expansion of the Beaker complex was driven to a substantial extent by migration. This genomic transformation is clearest in Britain owing to our densely sampled time transect. The arrival of people associated with the Beaker complex precipitated a demographic transformation in Britain, exemplified by the presence of individuals with large amounts of steppe-related ancestry after 2450 BC. We considered the possibility that an uneven geographic distribution of samples may have caused us to miss a major population that lacked steppe-derived ancestry after 2450 BC. However, our British Beaker and Bronze Age samples are dispersed geographically—extending from the southeastern peninsula of England to the Western Isles of Scotland—and come from a wide variety of funerary contexts (rivers, caves, pits, barrows, cists and flat graves) and diverse funerary traditions (single and multiple burials in variable states of anatomical articulation), which reduces the likelihood that our sampling missed major populations. We also considered the possibility that different burial practices between local and incoming populations (cremation versus inhumation) during the early stages of interaction could result in a sampling bias against local individuals. Although it is possible that such a sampling bias makes the ancestry transition appear more sudden than it in fact was, the long-term demographic effect was clearly substantial, as the pervasive steppe-related ancestry observed during the Beaker period, which was absent in the Neolithic period, persisted during the Bronze Age—and indeed remains predominant in

Britain today<sup>2</sup>. These results are notable in light of strontium and oxygen isotope analyses of British skeletons from the Beaker and Bronze Age periods<sup>30</sup>, which have provided no evidence for substantial mobility over individuals' lifetimes from locations with cooler climates or from places with geologies atypical of Britain. However, the isotope data are only sensitive to first-generation migrants and do not rule out movements from regions such as the lower Rhine area or from other geologically similar regions for which DNA sampling is still sparse. Further sampling of regions on the European continent may reveal additional candidate sources.

By analysing DNA data from ancient individuals, we have been able to provide constraints on the interpretations of the processes underlying cultural and social changes in Europe during the third millennium BC. Our results motivate further archaeological research to identify the changes in social organization, technology, subsistence, climate, population sizes<sup>31</sup> or pathogen exposure<sup>32,33</sup> that could have precipitated the demographic changes uncovered in this study.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 8 May 2017; accepted 4 January 2018.**

**Published online 21 February 2018.**

1. Czebreszuk, J. In *Ancient Europe, 8000 B.C. to A.D. 1000: An Encyclopedia of the Barbarian World* (eds Bogucki, P. I. & Crabtree, P. J.) 476–485 (Charles Scribner's Sons, 2004).
2. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
3. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
4. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
5. Czebreszuk, J. *Similar But Different. Bell Beakers in Europe* (Adam Mickiewicz Univ., 2004).
6. Cardoso, J. L. Absolute chronology of the Beaker phenomenon north of the Tagus estuary: demographic and social implications. *Trab. Prehist.* **71**, 56–75 (2014).
7. Jeunesse, C. The dogma of the Iberian origin of the Bell Beaker: attempting its deconstruction. *J. Neolit. Archaeol.* **16**, 158–166 (2015).
8. Fokkens, H. & Nicolis, F. *Background to Beakers. Inquiries into Regional Cultural Backgrounds of the Bell Beaker Complex* (Sidestone, 2012).
9. Linden, M. V. What linked the Bell Beakers in third millennium BC Europe? *Antiquity* **81**, 343–352 (2007).
10. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
11. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
12. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
13. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
14. Valverde, L. *et al.* New clues to the evolutionary history of the main European paternal lineage M269: dissection of the Y-SNP S116 in Atlantic Europe and Iberia. *Eur. J. Hum. Genet.* **24**, 437–441 (2016).
15. Gamba, C. *et al.* Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. *Mol. Ecol.* **21**, 45–56 (2012).
16. Günther, T. *et al.* Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl Acad. Sci. USA* **112**, 11917–11922 (2015).
17. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
18. Broushaki, F. *et al.* Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353**, 499–503 (2016).
19. Skoglund, P. *et al.* Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344**, 747–750 (2014).
20. Olalde, I. *et al.* A common genetic origin for early farmers from Mediterranean Cardial and central European LBK cultures. *Mol. Biol. Evol.* **32**, 3132–3142 (2015).
21. Mathieson, I. *et al.* The genomic history of southeastern Europe. *Nature* <https://doi.org/10.1038/nature25778> (2018).
22. Lipson, M. *et al.* Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature* **551**, 368–372 (2017).
23. Cassidy, L. M. *et al.* Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc. Natl Acad. Sci. USA* **113**, 368–373 (2016).
24. Sheridan, J. A. In *Landscapes in Transition* (eds Finlayson, B. & Warren, G.) 89–105 (Oxbow, 2010).

25. Burger, J., Kirchner, M., Bramanti, B., Haak, W. & Thomas, M. G. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc. Natl Acad. Sci. USA* **104**, 3736–3741 (2007).
26. Clarke, D. L. In *Glockenbecher Symposium, Oberried, 18–23 März 1974* (eds Lanting, J. N. & van der Waals, J. D.) 460–477 (Bussum, 1976).
27. Clark, G. The invasion hypothesis in British archaeology. *Antiquity* **40**, 172–189 (1966).
28. Brotherton, P. et al. Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nat. Commun.* **4**, 1764 (2013).
29. Desideri, J. *When Beakers Met Bell Beakers: an Analysis of Dental Remains (British Archaeological Reports International Series 2292)* (Archaeopress, 2011).
30. Parker Pearson, M. et al. Beaker people in Britain: migration, mobility and diet. *Antiquity* **90**, 620–637 (2016).
31. Shennan, S. et al. Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nat. Commun.* **4**, 2486 (2013).
32. Valtueña, A. A. et al. The Stone Age plague and its persistence in Eurasia. *Curr. Biol.* **27**, 3683–3691 (2017).
33. Rasmussen, S. et al. Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* **163**, 571–582 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank D. Anthony, J. Koch, I. Mathieson and C. Renfrew for comments; A. Cooper for support from the Australian Centre for Ancient DNA; the Bristol Radiocarbon Accelerator Mass Spectrometry Facility (BRAMS); A. C. Sousa, A. M. Cölliga, L. Loe, C. Roth, E. Carmona Ballesteros, M. Kunst, S.-A. Coupard, M. Giesen, T. Lord, M. Green, A. Chamberlain and G. Drinkall for assistance with samples; E. Willerslev for supporting several co-authors at the Centre for GeoGenetics; the Museo Arqueológico Regional de la Comunidad de Madrid, the Hunterian Museum, University of Glasgow, the Orkney Museum, the Museu Municipal de Torres Vedras, the Great North Museum: Hancock, the Society of Antiquaries of Newcastle upon Tyne, the Sunderland Museum, the National Museum of Wales, the Duckworth Laboratory, the Wiltshire Museum, the Wells Museum, the Brighton Museum, the Somerset Heritage Museum and the Museum of London for facilitating sample collection. Support for this project was provided by Czech Academy of Sciences grant RVO:67985912; by the Momentum Mobility Research Group of the Hungarian Academy of Sciences; by the Wellcome Trust (100713/Z/12/Z); by Irish Research Council grant GOIPG/2013/36 to D.F.; by the Heidelberg Academy of Sciences (WIN project ‘Times of Upheaval’) to P.W.S., J.K. and A.Mi.; by the Swedish Foundation for Humanities and Social Sciences grant M16-0455:1 to K.Kr.; by the National Science Centre, Poland grant DEC-2013/10/E/HS3/00141 to M.Fu.; by Obra Social La Caixa and by a Spanish MINECO grant BFU2015-64699-P to C.L.-F.; by a Spanish MINECO grant HAR2016-77600-P to C.L., P.R. and C.Bi.; by the NSF Archaeometry program BCS-1460369 to D.J.K.; by the NFS Archaeology program BCS-1725067 to D.J.K. and T.Ha.; and by an Allen Discovery Center grant from the Paul Allen Foundation, US National Science Foundation HOMINID grant BCS-1032255, US National Institutes of Health grant GM100233, and the Howard Hughes Medical Institute to D.R.

**Author Contributions** S.B., M.E.A., N.R., A.S.-N., A.Mi., N.B., M.Fe., E.Ha., M.Mi., J.O., K.S., O.C., D.K., F.C., R.Pi., J.K., W.H., I.B. and D.R. performed or supervised laboratory work. G.T.C. and D.J.K. undertook the radiocarbon dating of a large fraction of samples. I.A., K.Kr., A.B., K.W.A., A.A.F., E.B., M.B.-B., D.B., C.Bi., J.V.M., R.M.G., C.Bo., L.Bo., T.A., L.Bü., S.C., L.C.N., O.E.C., G.T.C., B.C., A.D., K.E.D., N.D., M.E., C.E., M.K., J.F.F., H.F., C.F., M.G., R.G.P., M.H.-U., E.Had., G.H., N.J., T.K., K.Ma., S.P., O.L., A.L., C.H.M., V.G.O., A.B.R., J.L.M., T.M., J.I.M., K.Mc., B.G.M., A.Mo., G.K., V.K., A.C., R.Pa., A.E., K.Kö., T.Ha., T.S., J.Da., Z.B., M.H., P.V., M.D., F.B., R.F.F., A.-M.H.-C., S.T., E.C., L.L., A.V., A.Z., C.W., G.D., E.G.-D., B.N., M.Br., M.L.P., R.M., J.De., M.Be., M.Fu., A.H., M.Ma., A.R., S.L., I.S., K.T.L., J.L.C., C.L., M.P.P., P.W., T.D.P., P.P., P.-J.R., P.R., R.R., M.A.R.G., A.Sc., J.S., A.M.S., V.S., L.V., J.Z., D.C., T.Hi., V.H., A.Sh., K.-G.S., P.W.S., R.Pi., J.K., W.H., I.B., C.L.-F. and D.R. assembled archaeological material. I.O., S.M., T.B., A.Mi., E.A., M.Li., I.L., N.P., Y.D., Z.F., D.F., D.J.K., P.d.K., T.K.H., M.G.T. and D.R. analysed data. I.O., C.L.-F. and D.R. wrote the manuscript with input from all co-authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to I.O. ([inigo\\_olalde@hms.harvard.edu](mailto:inigo_olalde@hms.harvard.edu)) or D.R. ([reich@genetics.med.harvard.edu](mailto:reich@genetics.med.harvard.edu)).

**Reviewer Information** Nature thanks C. Renfrew, E. Rhodes, M. Richards and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Iñigo Olalde<sup>1</sup>, Selina Brace<sup>2</sup>, Morten E. Allentoft<sup>3</sup>, Ian Armit<sup>4</sup>§, Kristian Kristiansen<sup>5</sup>, Thomas Booth<sup>6</sup>, Nadin Rohland<sup>1</sup>, Swapan Mallick<sup>1,6,7</sup>, Anna Szécsényi-Nagy<sup>8</sup>, Alissa Mittnik<sup>9,10</sup>, Eveline Altena<sup>11</sup>, Mark Lipson<sup>1</sup>, Iosif Lazaridis<sup>1,6</sup>, Thomas K. Harper<sup>12</sup>, Nick Patterson<sup>6</sup>, Nasreen Broomandkhosbacht<sup>1,7</sup>, Yoan Diekmann<sup>13</sup>, Zuzana Faltyskova<sup>13</sup>, Daniel Fernandes<sup>14,15,16</sup>, Matthew Ferry<sup>1,7</sup>, Eadaoin Harney<sup>1</sup>, Peter de Knijff<sup>1</sup>,

Megan Michel<sup>1,7</sup>, Jonas Oppenheimer<sup>1,7</sup>, Kristin Stewardson<sup>1,7</sup>, Alistair Barclay<sup>1,7</sup>, Kurt Werner Alt<sup>18,19,20</sup>, Corina Liesau<sup>21</sup>, Patricia Ríos<sup>21</sup>, Concepción Blasco<sup>21</sup>, Jorge Vega Miguel<sup>22</sup>, Roberto Menduina García<sup>22</sup>, Azucena Avilés Fernández<sup>23</sup>, Eszter Bánffy<sup>24,25</sup>, Maria Bernabò-Brea<sup>26</sup>, David Billon<sup>27</sup>, Clive Bonsall<sup>28</sup>, Laura Bonsall<sup>29</sup>, Tim Allen<sup>30</sup>, Lindsey Büster<sup>4</sup>, Sophie Carver<sup>31</sup>, Laura Castells Navarro<sup>4</sup>, Oliver E. Craig<sup>32</sup>, Gordon T. Cook<sup>33</sup>, Barry Cunliffe<sup>34</sup>, Anthony Denaire<sup>35</sup>, Kirsten Egging Dinwiddie<sup>17</sup>, Natasha Dowdell<sup>36</sup>, Michal Ernee<sup>37</sup>, Christopher Evans<sup>38</sup>, Milan Kuchařik<sup>39</sup>, Joan Francès Farré<sup>40</sup>, Chris Fowler<sup>41</sup>, Michiel Hazenbeek<sup>42</sup>, Rafael Garrido Pena<sup>21</sup>, Maria Haber-Uriarte<sup>23</sup>, Elżbieta Haduch<sup>43</sup>, Gill Hey<sup>30</sup>, Nick Jowett<sup>44</sup>, Timothy Knowles<sup>45</sup>, Ken Massy<sup>46</sup>, Saskia Pfrenkle<sup>9</sup>, Philippe Lefranc<sup>47</sup>, Olivier Lemerrier<sup>48</sup>, Arnaud Lefebvre<sup>49,50</sup>, César Heras Martínez<sup>51,52,53</sup>, Virginia Galera Olmo<sup>52,53</sup>, Ana Bastida Ramírez<sup>51</sup>, Joaquín Lomba Maurandi<sup>23</sup>, Tona Majó<sup>54</sup>, Jacqueline I. McKinley<sup>17</sup>, Kathleen McSweeney<sup>28</sup>, Balázs Gusztáv Mende<sup>8</sup>, Alessandra Mod<sup>55</sup>, Gabriella Kulcsár<sup>24</sup>, Viktória Kiss<sup>24</sup>, András Czene<sup>56</sup>, Róbert Patay<sup>57</sup>, Anna Endrődi<sup>58</sup>, Kitti Köhler<sup>24</sup>, Tamás Hajdu<sup>59,60</sup>, Tamás Szeniczey<sup>59</sup>, János Dani<sup>61</sup>, Zsolt Bernert<sup>60</sup>, Maya Hoole<sup>62</sup>, Olivia Cheronet<sup>14,15</sup>, Denise Keating<sup>63</sup>, Petr Velemínský<sup>64</sup>, Miroslav Dobes<sup>37</sup>, Francesca Candilio<sup>65,66,67</sup>, Fraser Brown<sup>30</sup>, Raúl Flores Fernández<sup>68</sup>, Ana-Mercedes Herrero-Corral<sup>69</sup>, Sebastiano Tusa<sup>70</sup>, Emiliano Carnieri<sup>71</sup>, Luigi Lentini<sup>72</sup>, Antonella Valenti<sup>73</sup>, Alessandro Zanini<sup>74</sup>, Clive Waddington<sup>75</sup>, Germán Delibes<sup>76</sup>, Elisa Guerra-Doce<sup>76</sup>, Benjamin Neil<sup>38</sup>, Marcus Brittain<sup>38</sup>, Mike Luke<sup>77</sup>, Richard Mortimer<sup>36</sup>, Jocelyne Desideri<sup>78</sup>, Marie Besse<sup>78</sup>, Günter Brückner<sup>79</sup>, Mirosław Furmanek<sup>80</sup>, Agata Haliusko<sup>80</sup>, Maksym Mackiewicz<sup>80</sup>, Artur Rapiński<sup>81</sup>, Stephany Leach<sup>82</sup>, Ignacio Soriano<sup>83</sup>, Katina T. Lillios<sup>84</sup>, João Luís Cardoso<sup>85,86</sup>, Michael Parker Pearson<sup>87</sup>, Piotr Włodarczyk<sup>88</sup>, T. Douglas Price<sup>89</sup>, Pilar Prieto<sup>90</sup>, Pierre-Jérôme Rey<sup>91</sup>, Roberto Risch<sup>83</sup>, Manuel A. Rojo Guerra<sup>82</sup>, Aurore Schmitt<sup>93</sup>, Joël Serrallongue<sup>94</sup>, Ana Maria Silva<sup>95</sup>, Václav Smrčka<sup>96</sup>, Luc Vergnaud<sup>97</sup>, João Zilhão<sup>85,98,99</sup>, David Caramelli<sup>55</sup>, Thomas Higham<sup>100</sup>, Mark G. Thomas<sup>13</sup>, Douglas J. Kennet<sup>101</sup>, Harry Fokkens<sup>102</sup>, Volker Heyd<sup>101,103</sup>, Alison Sheridan<sup>104</sup>, Karl-Göran Sjögren<sup>5</sup>, Philipp W. Stockhammer<sup>46,105</sup>, Johannes Krause<sup>105</sup>, Ron Pinhasi<sup>14,15</sup>§, Wolfgang Haak<sup>105,106</sup>§, Ian Barnes<sup>2</sup>§, Carles Lalueza-Fox<sup>107</sup>§ & David Reich<sup>1,6,7</sup>§

<sup>1</sup>Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

<sup>2</sup>Department of Earth Sciences, Natural History Museum, London SW7 5BD, UK. <sup>3</sup>Centre for GeoGenetics, Natural History Museum, University of Copenhagen, Copenhagen 1350, Denmark.

<sup>4</sup>School of Archaeological and Forensic Sciences, University of Bradford, Bradford BD7 1DP, UK.

<sup>5</sup>University of Gothenburg, Gothenburg 405 30, Sweden. <sup>6</sup>Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. <sup>7</sup>Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>8</sup>Laboratory of Archaeogenetics, Institute of Archaeology, Research Centre for the Humanities, Hungarian Academy of Sciences, Budapest 1097, Hungary. <sup>9</sup>Institute for Archaeological Sciences, Archaeo- and Palaeogenetics, University of Tübingen, Tübingen 72070, Germany. <sup>10</sup>Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena 07745, Germany. <sup>11</sup>Department of Human Genetics, Leiden University Medical Center, Leiden 2333 ZC, The Netherlands. <sup>12</sup>Department of Anthropology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.

<sup>13</sup>Research Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK. <sup>14</sup>Earth Institute, University College Dublin, Dublin 4, Ireland.

<sup>15</sup>Department of Anthropology, University of Vienna, Vienna 1090, Austria. <sup>16</sup>Research Center for Anthropology and Health, Department of Life Science, University of Coimbra, Coimbra 3000-456, Portugal. <sup>17</sup>Wessex Archaeology, Salisbury SP4 6EB, UK. <sup>18</sup>Center of Natural and Cultural History of Man, Danube Private University, Krems 3500, Austria. <sup>19</sup>Department of Biomedical Engineering, Basel University, Basel 4123, Switzerland. <sup>20</sup>Integrative Prehistory and Archaeological Science, Basel University, Basel, Switzerland. <sup>21</sup>Departamento de Prehistoria y Arqueología, Universidad Autónoma de Madrid, Madrid 28049, Spain. <sup>22</sup>ARGE S.L., Madrid 28011, Spain. <sup>23</sup>Área de Prehistoria, Universidad de Murcia, Murcia 30001, Spain. <sup>24</sup>Institute of Archaeology, Research Centre for the Humanities, Hungarian Academy of Sciences, Budapest 1097, Hungary. <sup>25</sup>Romano-Germanic Commission, German Archaeological Institute, Frankfurt am Main 60325, Germany. <sup>26</sup>Museo Archeologico Nazionale di Parma, Parma 43100, Italy.

<sup>27</sup>INRAP, Institut National de Recherches Archéologiques Préventives, Buffard 25440, France.

<sup>28</sup>School of History, Classics and Archaeology, University of Edinburgh, Edinburgh EH8 9AG, UK. <sup>29</sup>10 Merchiston Gardens, Edinburgh EH10 5DD, UK. <sup>30</sup>Oxford Archaeology, Oxford OX2 0ES, UK. <sup>31</sup>Department of Archaeology and Anthropology, University of Bristol, Bristol BS8 1UU, UK. <sup>32</sup>BioArch, Department of Archaeology, University of York, York YO10 5DD, UK. <sup>33</sup>Scottish Universities Environmental Research Centre, East Kilbride G75 0QF, UK. <sup>34</sup>Institute of Archaeology, University of Oxford, Oxford OX1 2PG, UK. <sup>35</sup>University of Burgundy, Dijon 21000, France. <sup>36</sup>Oxford Archaeology East, Cambridge CB23 8SQ, UK. <sup>37</sup>Institute of Archaeology, Czech Academy of Sciences, Prague 118 01, Czech Republic. <sup>38</sup>Cambridge Archaeological Unit, Department of Archaeology, University of Cambridge, Cambridge CB3 0DT, UK. <sup>39</sup>Labrys o.p.s., Prague 198 00, Czech Republic. <sup>40</sup>Museu i Poblat Ibèric de Ca n'Oliver, Cerdanyola del Vallès 08290, Spain. <sup>41</sup>School of History, Classics & Archaeology, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. <sup>42</sup>INRAP, Institut National de Recherches Archéologiques Préventives, Nice 06300, France. <sup>43</sup>Institute of Zoology and Biomedical Research, Jagiellonian University, Kraków 31-007, Poland. <sup>44</sup>Great Orme Mines, Great Orme, Llandudno LL30 2XG, UK. <sup>45</sup>Bristol Radiocarbon Accelerator Mass Spectrometry Facility, University of Bristol, Bristol BS8 1UU, UK.

<sup>46</sup>Institut für Vor- und Frühgeschichtliche Archäologie und Provinzialrömische Archäologie, Ludwig-Maximilians-Universität München, Munich 80539, Germany. <sup>47</sup>INRAP, Institut National de Recherches Archéologiques Préventives, Strasbourg 67100, France. <sup>48</sup>Université Paul-Valéry - Montpellier 3, UMR 5140 ASM, Montpellier 34199, France. <sup>49</sup>INRAP, Institut National de

Recherches Archéologiques Préventives, Metz 57063, France. <sup>50</sup>UMR 5199, Pacea, équipe A3P, Université de Bordeaux, Talence 33400, France. <sup>51</sup>TRÉBEDE, Patrimonio y Cultura SL, Torres de la Alameda 28813, Spain. <sup>52</sup>Departamento de Ciencias de la Vida, Universidad de Alcalá, Alcalá de Henares 28801, Spain. <sup>53</sup>Instituto Universitario de Investigación en Ciencias Policiales (IUICP), Alcalá de Henares 28801, Spain. <sup>54</sup>Archaeom, Departament de Prehistòria, Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain. <sup>55</sup>Department of Biology, University of Florence, Florence 50121, Italy. <sup>56</sup>Salisbury Ltd, Budaörs 2040, Hungary. <sup>57</sup>Ferenczy Museum Center, Szentendre 2100, Hungary. <sup>58</sup>Budapest History Museum, Budapest 1014, Hungary. <sup>59</sup>Department of Biological Anthropology, Eötvös Loránd University, Budapest 1117, Hungary. <sup>60</sup>Hungarian Natural History Museum, Budapest 1083, Hungary. <sup>61</sup>Déri Museum, Debrecen 4026, Hungary. <sup>62</sup>Historic Environment Scotland, Edinburgh EH9 1SH, UK. <sup>63</sup>Humanities Institute, University College Dublin, Dublin 4, Ireland. <sup>64</sup>Department of Anthropology, National Museum, Prague 115 79, Czech Republic. <sup>65</sup>Soprintendenza Archeologia belle arti e paesaggio per la città metropolitana di Cagliari e per le province di Oristano e Sud Sardegna, Cagliari 9124, Italy. <sup>66</sup>Physical Anthropology Section, University of Philadelphia Museum of Archaeology and Anthropology, Philadelphia, Pennsylvania 19104, USA. <sup>67</sup>Department of Environmental Biology, Sapienza University of Rome, Rome 00185, Italy. <sup>68</sup>46 Ciudad Real Street, Parla 28982, Spain. <sup>69</sup>Departamento de Prehistoria, Universidad Complutense de Madrid, Madrid 28040, Spain. <sup>70</sup>Soprintendenza del Mare, Palermo 90133, Italy. <sup>71</sup>Facoltà di Lettere e Filosofia, Università di Palermo, Palermo 90133, Italy. <sup>72</sup>Soprintendenza per i beni culturali e ambientali di Trapani, Trapani 91100, Italy. <sup>73</sup>Prima Archeologia del Mediterraneo, Partanna 91028, Italy. <sup>74</sup>Università degli Studi di Palermo, Agrigento 92100, Italy. <sup>75</sup>Archaeological Research Services Ltd, Bakewell DE45 1HB, UK. <sup>76</sup>Departamento de Prehistoria, Facultad de Filosofía y Letras, Universidad de Valladolid, Valladolid 47011, Spain. <sup>77</sup>Albion Archaeology, Bedford MK42 0AS, UK. <sup>78</sup>Laboratory of Prehistoric Archaeology and Anthropology, Department F.-A. Forel for Environmental and Aquatic Sciences, University of Geneva, Geneva 4, Switzerland. <sup>79</sup>General Department of Cultural Heritage Rhineland Palatinate, Department of Archaeology, Mainz 55116, Germany. <sup>80</sup>Institute of Archaeology, University of Wrocław, Wrocław 50-137, Poland. <sup>81</sup>Institute of

Archaeology, Silesian University in Opava, Opava 746 01, Czech Republic. <sup>82</sup>Department of Archaeology, University of Exeter, Exeter EX4 4QE, UK. <sup>83</sup>Departament de Prehistòria, Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain. <sup>84</sup>Department of Anthropology, University of Iowa, Iowa City, Iowa 52240, USA. <sup>85</sup>Centro de Arqueologia, Universidade de Lisboa, Lisboa 1600-214, Portugal. <sup>86</sup>Universidade Aberta, Lisboa 1269-001, Portugal. <sup>87</sup>Institute of Archaeology, University College London, London WC1H 0PY, UK. <sup>88</sup>Institute of Archaeology and Ethnology, Polish Academy of Sciences, Kraków 31-016, Poland. <sup>89</sup>Laboratory for Archaeological Chemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. <sup>90</sup>University of Santiago de Compostela, Santiago de Compostela 15782, Spain. <sup>91</sup>UMR 5204 Laboratoire Edytem, Université Savoie Mont Blanc, Chambéry 73376, France. <sup>92</sup>Department of Prehistory and Archaeology, Faculty of Philosophy and Letters, Valladolid University, Valladolid 47011, Spain. <sup>93</sup>UMR 7268 ADES, CNRS, Aix-Marseille Univ, EFS, Faculté de médecine Nord, Marseille 13015, France. <sup>94</sup>Service archéologique, Conseil Général de la Haute-Savoie, Annecy 74000, France. <sup>95</sup>Laboratory of Prehistory, Research Center for Anthropology and Health, Department of Life Science, University of Coimbra, Coimbra 3000-456, Portugal. <sup>96</sup>Institute for History of Medicine and Foreign Languages, First Faculty of Medicine, Charles University, Prague 121 08, Czech Republic. <sup>97</sup>ANTEA Bureau d'étude en Archéologie, Habsheim 68440, France. <sup>98</sup>Institució Catalana de Recerca i Estudis Avançats, Barcelona 08010, Spain. <sup>99</sup>Departament d'Història i Arqueologia, Universitat de Barcelona, Barcelona 08001, Spain. <sup>100</sup>Oxford Radiocarbon Accelerator Unit, RLAHA, University of Oxford, Oxford OX1 3QY, UK. <sup>101</sup>Department of Anthropology & Institute for Energy and the Environment, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. <sup>102</sup>Faculty of Archaeology, Leiden University, 2333 CC Leiden, The Netherlands. <sup>103</sup>Department of Philosophy, History, Culture and Art Studies, Section of Archaeology, University of Helsinki, Helsinki 00014, Finland. <sup>104</sup>National Museums Scotland, Edinburgh EH1 1JF, UK. <sup>105</sup>Max Planck Institute for the Science of Human History, Jena 07745, Germany. <sup>106</sup>Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide 5005, South Australia, Australia. <sup>107</sup>Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona 08003, Spain.

§These authors jointly supervised this work.



## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Ancient DNA analysis.** We screened skeletal samples for DNA preservation in dedicated clean rooms. We extracted DNA<sup>34–36</sup> and prepared barcoded next-generation sequencing libraries, the majority of which were treated with uracil-DNA glycosylase (UDG) to greatly reduce the damage (except at the terminal nucleotide) that is characteristic of ancient DNA<sup>37,38</sup> (Supplementary Information section 4). We initially enriched libraries for sequences overlapping the mitochondrial genome<sup>39</sup> and approximately 3,000 nuclear SNPs, using synthesized baits (CustomArray) that we PCR-amplified. We sequenced the enriched material on an Illumina NextSeq instrument with  $2 \times 76$  cycles, and  $2 \times 7$  cycles to read out the two indices<sup>40</sup>. We merged read pairs with the expected barcodes that overlapped by at least 15 bases, mapped the merged sequences to the human reference genome hg19 and to the reconstructed mitochondrial DNA consensus sequence<sup>41</sup> using the 'samse' command in bwa v.0.6.1<sup>42</sup>, and then removed duplicated sequences. We evaluated DNA authenticity by estimating the rate of mismatching to the consensus mitochondrial sequence<sup>43</sup>, and also by requiring that the rate of damage at the terminal nucleotide was at least 3% for UDG-treated libraries<sup>43</sup> and 10% for non-UDG-treated libraries<sup>44</sup>.

For libraries that appeared promising after screening, we enriched in two consecutive rounds for sequences overlapping 1,233,013 SNPs ('1,240k SNP capture')<sup>2,10</sup> and sequenced  $2 \times 76$  cycles and  $2 \times 7$  cycles on an Illumina NextSeq500 instrument. We bioinformatically processed the data in the same way as for the mitochondrial capture data, except that this time we mapped only to hg19 and merged the data from different libraries of the same individual. We further evaluated authenticity by looking at the ratio of X-to-Y chromosome reads and estimating X-chromosome contamination in males based on the rate of heterozygosity<sup>45</sup>. Samples with evidence of contamination were either filtered out or restricted to sequences with terminal cytosine deamination in order to remove sequences that derived from modern contaminants. Finally, we filtered out samples with fewer than 10,000 targeted SNPs covered at least once and samples that were first-degree relatives of others in the dataset (keeping the sample with the larger number of covered SNPs) (Supplementary Table 1) from our genome-wide analysis dataset. **Mitochondrial haplogroup determination.** We used the mitochondrial capture .bam files to determine the mitochondrial haplogroup of each sample with new data, restricting our analysis to sequences with MAPQ  $\geq 30$  and base quality  $\geq 30$ . First, we constructed a consensus sequence with samtools and bcftools<sup>46</sup>, using a majority rule and requiring a minimum coverage of two. We called haplogroups with HaploGrep<sup>247</sup> based on phylotree<sup>48</sup> (mtDNA tree build 17 (accessed 18 February 2016)). Mutational differences, compared to the revised Cambridge Reference Sequence (GenBank reference sequence: NC\_012920.1) and corresponding haplogroups, can be viewed in Supplementary Table 2. We computed haplogroup frequencies for relevant ancient populations (Supplementary Table 3) after removing close relatives with the same mtDNA.

**Y-chromosome analysis.** We determined Y-chromosome haplogroups for both new and published samples (Supplementary Information section 5). We made use of the sequences mapping to 1,240k Y-chromosome targets, restricting our analysis to sequences with mapping quality  $\geq 30$  and bases with quality  $\geq 30$ . We called haplogroups by determining the most derived mutation for each sample, using the nomenclature of the International Society of Genetic Genealogy (<http://www.isogg.org>) version 11.110 (accessed 21 April 2016). Haplogroups and their supporting derived mutations can be viewed in Supplementary Table 4.

**Merging newly generated data with published data.** We assembled two datasets for genome-wide analyses. The first dataset is HO, which includes 2,572 present-day individuals from worldwide populations genotyped on the Human Origins Array<sup>11,12,49</sup> and 683 ancient individuals. The ancient set includes 211 Beaker-complex-associated individuals (195 newly reported, 7 with shotgun data<sup>3</sup> for which we generated 1,240k capture data and 9 that had previously been published<sup>3,4</sup>), 68 newly reported individuals from relevant ancient populations and 298 individuals that had previously been published<sup>12,18,19,21–23,50–57</sup> (Supplementary Table 1). We kept 591,642 autosomal SNPs after intersecting autosomal SNPs in the 1,240k capture with the analysis set of 594,924 SNPs from a previous publication<sup>11</sup>. The second dataset is HOIII, which includes the same set of ancient samples and 300 present-day individuals from 142 populations sequenced to high coverage as part of the Simons Genome Diversity Project<sup>13</sup>. For this dataset, we used 1,054,671 autosomal SNPs, excluding SNPs of the 1,240k array located on sex chromosomes or with known functional effects.

For each individual, we represented the allele at each SNP by randomly sampling one sequence and discarding the first and the last two nucleotides of each sequence.

**Abbreviations.** We have used the following abbreviations in population labels: E, Early; M, Middle; L, Late; N, Neolithic; CA, Copper Age; BA, Bronze Age; BC, Beaker complex; N\_Iberia, northern Iberia; C\_Iberia, central Iberia; SE\_Iberia, southeast Iberia; and SW\_Iberia, southwest Iberia.

**Principal component analysis.** We carried out principal component analysis on the HO dataset using the 'smartpca' program in EIGENSOFT<sup>58</sup>. We computed principal components on 990 present-day west Eurasians and projected ancient individuals using Isqproject: YES and shrinkmode: YES.

**ADMIXTURE analysis.** We performed model-based clustering analysis using ADMIXTURE<sup>17</sup> on the HO reference dataset, which included 2,572 present-day individuals from worldwide populations and the ancient individuals. First, we carried out linkage disequilibrium pruning on the dataset using PLINK<sup>59</sup> with the flag --indep-pairwise 200 25 0.4, leaving 306,393 SNPs. We ran ADMIXTURE with the cross validation (--cv) flag specifying from  $K=2$  to  $K=20$  clusters, with 20 replicates for each value of  $K$ . For each value of  $K$ , the replicate with highest log likelihood was kept. In Extended Data Fig. 3b we show the cluster assignments at  $K=8$  of newly reported individuals and other relevant ancient samples for comparison. We chose this value of  $K$  as it was the lowest one for which components of ancestry related both to Iranian Neolithic farmers and European Mesolithic hunter-gatherers were maximized.

**f-statistics.** We computed  $f$ -statistics on the HOIII dataset using ADMIXTOOLS<sup>49</sup> with default parameters (Supplementary Information section 6). We used qpDstat with f4mode: Yes for  $f_4$ -statistics and qp3Pop for outgroup  $f_3$ -statistics. We computed standard errors using a weighted block jackknife<sup>60</sup> over 5-Mb blocks.

**Inference of mixture proportions.** We estimated ancestry proportions on the HOIII dataset using qpAdm<sup>2</sup> and a basic set of nine outgroups: Mota, Ust\_Ishim, MA1, Villabruna, Mbuti, Papuan, Onge, Han and Karitiana. For some analyses (Supplementary Information section 8) we added additional outgroups to this basic set.

**Admixture graph modelling.** We modelled the relationships between populations in an Admixture Graph framework with the software qpGraph in ADMIXTOOLS<sup>49</sup>, using the HOIII dataset and Mbuti as an outgroup (Supplementary Information section 7).

**Allele frequency estimation from read counts.** We used allele counts at each SNP to perform maximum likelihood estimations of allele frequencies in ancient populations as in ref. 4. In Extended Data Fig. 7, we show derived allele frequency estimates at three SNPs of functional importance for different ancient populations.

**Data availability.** All 1,240k and mitochondrial capture sequencing data are available from the European Nucleotide Archive, accession number PRJEB23635. The genotype dataset is available from the Reich Laboratory website at <https://reich.hms.harvard.edu/datasets>.

34. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110**, 15758–15763 (2013).
35. Damgaard, P. B. *et al.* Improving access to endogenous DNA in ancient bones and teeth. *Sci. Rep.* **5**, 11184 (2015).
36. Korlević, P. *et al.* Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *Biotechniques* **59**, 87–93 (2015).
37. Rohland, N., Harney, E., Mallick, S., Nordenfellt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B* **370**, 20130624 (2015).
38. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
39. Maricic, T., Whitten, M. & Pääbo, S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* **5**, e14004 (2010).
40. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).
41. Behar, D. M. *et al.* A 'Copernican' reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90**, 675–684 (2012).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013).
44. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Pääbo, S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* **7**, e34131 (2012).
45. Korneliusson, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
46. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
47. Weissensteiner, H. *et al.* HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).

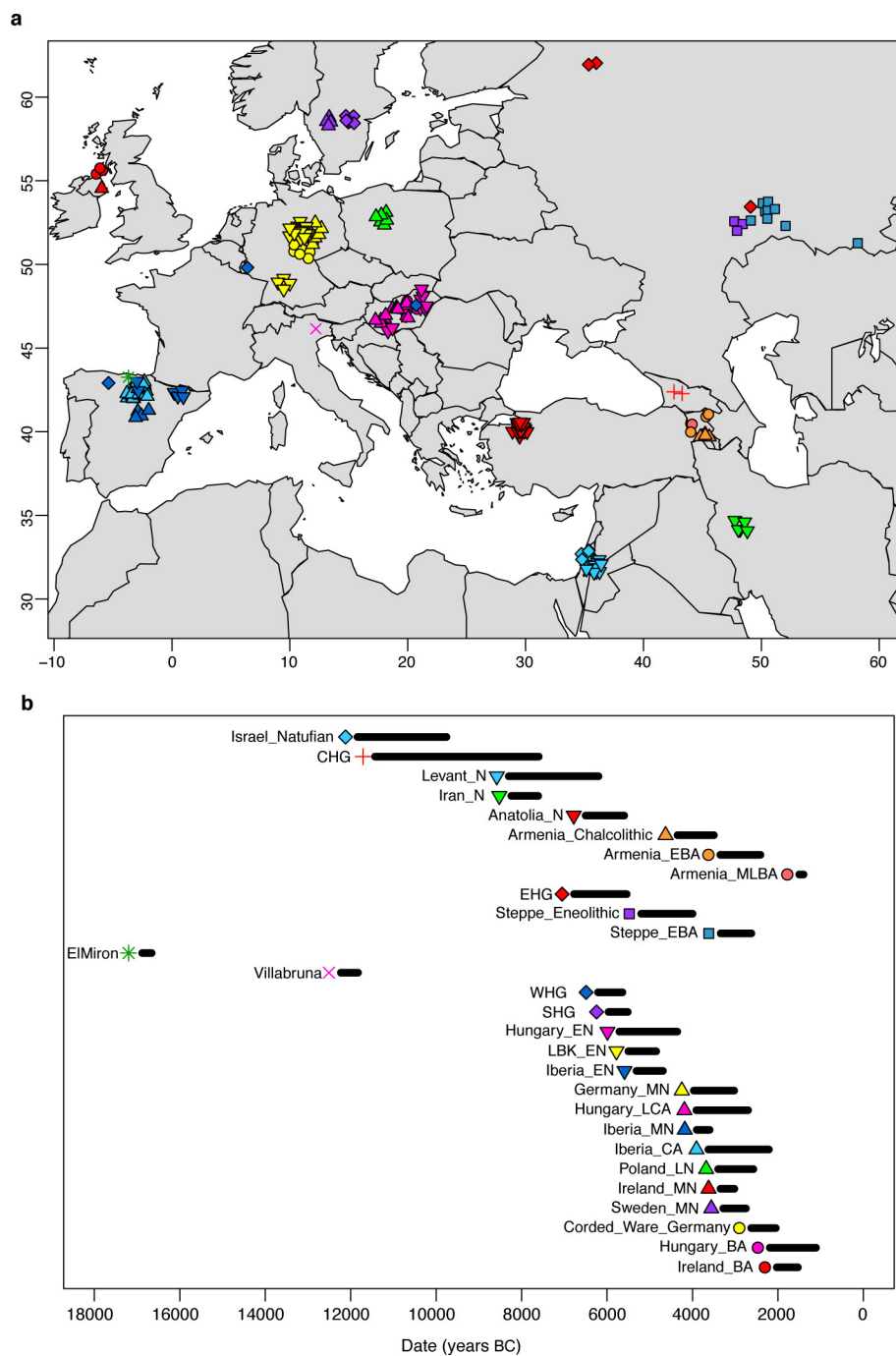
48. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–E394 (2009).
49. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
50. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).
51. Omrak, A. *et al.* Genomic evidence establishes Anatolia as the source of the European Neolithic gene pool. *Curr. Biol.* **26**, 270–275 (2016).
52. Gallego Llorente, M. *et al.* Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science* **350**, 820–822 (2015).
53. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).
54. Kiliç, G. M. *et al.* The demographic development of the first farmers in Anatolia. *Curr. Biol.* **26**, 2659–2666 (2016).
55. Gallego-Llorente, M. *et al.* The genetics of an early Neolithic pastoralist from the Zagros, Iran. *Sci. Rep.* **6**, 31326 (2016).
56. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–228 (2014).
57. Hofmanová, Z. *et al.* Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl Acad. Sci. USA* **113**, 6886–6891 (2016).
58. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
59. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
60. Busing, F. M. T. A., Meijer, E. & Van Der Leeden, R. Delete-m jackknife for unequal m. *Stat. Comput.* **9**, 3–8 (1999).
61. Rojo-Guerra, M. Á., Kunst, M., Garrido-Pena, R. & García-Martínez de Lagrán, I. & Morán-Dauchez, G. *Un desafío a la eternidad. Tumbas monumentales del Valle de Ambrona (Memorias Arqueología en Castilla y León 14)* (Junta de Castilla y León, 2005).
62. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014).



**Extended Data Figure 1 | Beaker-complex artefacts.** **a**, 'All-Over-Cord' Beaker from Bathgate, West Lothian, Scotland. Photograph: © National Museums Scotland. **b**, Beaker-complex grave goods from La Sima III

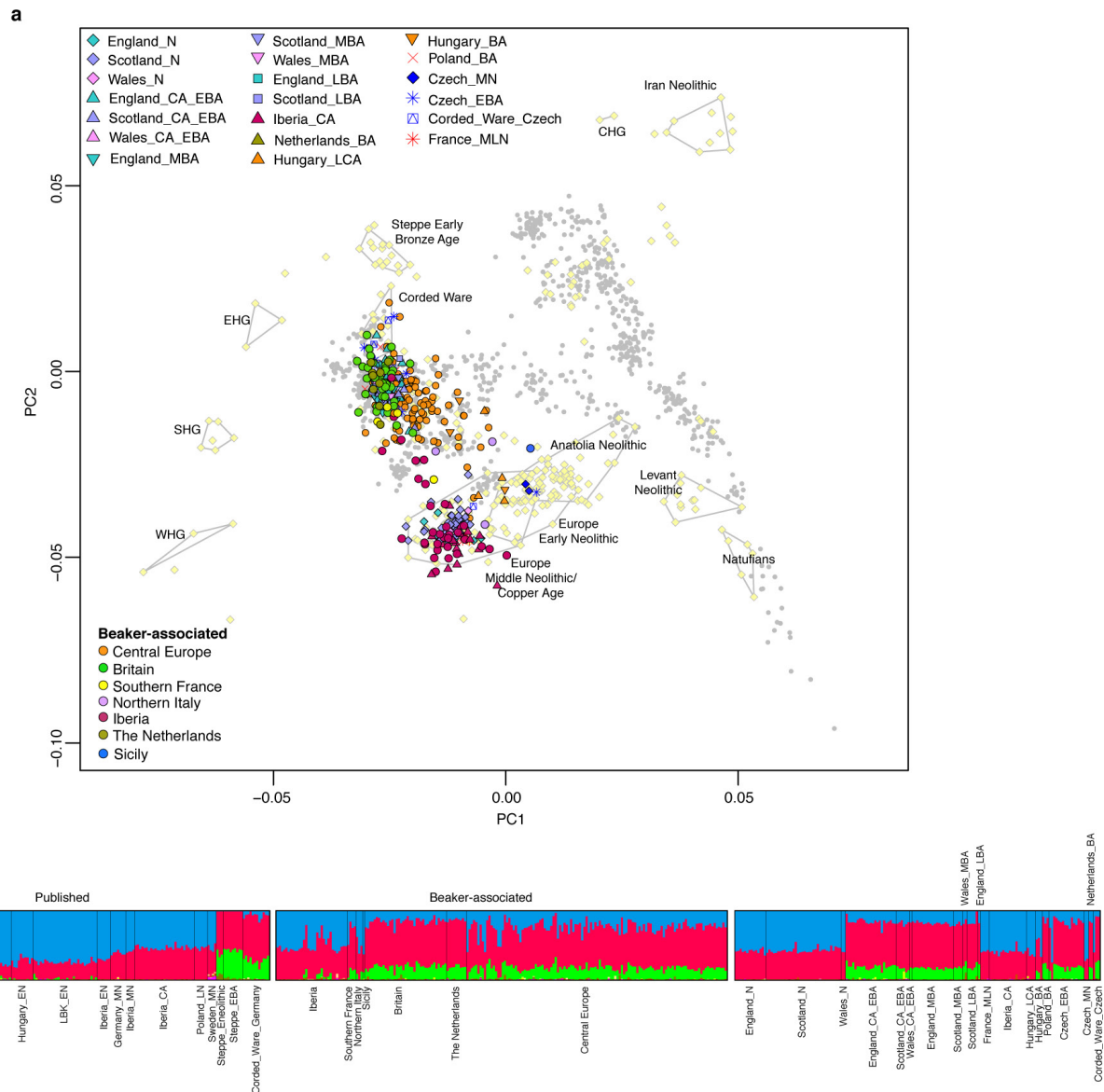
barrow, Soria, Spain<sup>61</sup>. The set includes Beaker pots of the so-called 'Maritime style'. Photograph: Junta de Castilla y León, Archivo Museo Numantino, Alejandro Plaza.





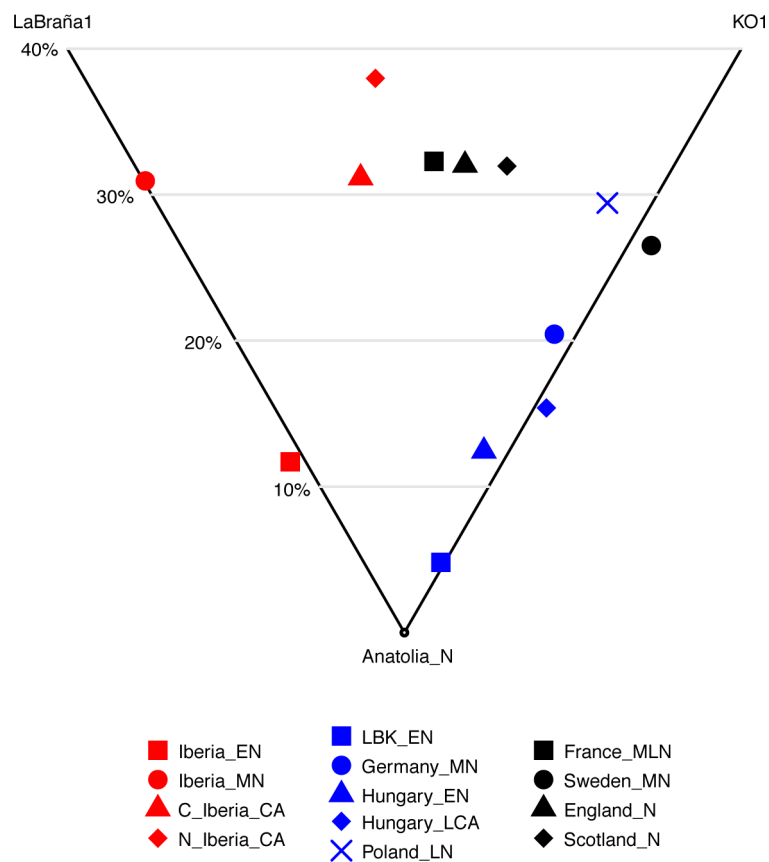
**Extended Data Figure 2 | Ancient individuals with previously published genome-wide data used in this study. a, Sampling locations. b, Time ranges.** WHG, western hunter-gatherers; EHG, eastern hunter-gatherers;

SHG, Scandinavian hunter-gatherers; CHG, Caucasus hunter-gatherers; E, Early; M, Middle; L, Late; N, Neolithic; CA, Copper Age; and BA, Bronze Age. Map data from the R package 'maps'.



**Extended Data Figure 3 | Population structure. a**, Principal component analysis of 990 present-day west Eurasian individuals (grey dots), with previously published (pale yellow) and new ancient samples projected onto the first two principal components. **b**, ADMIXTURE clustering analysis

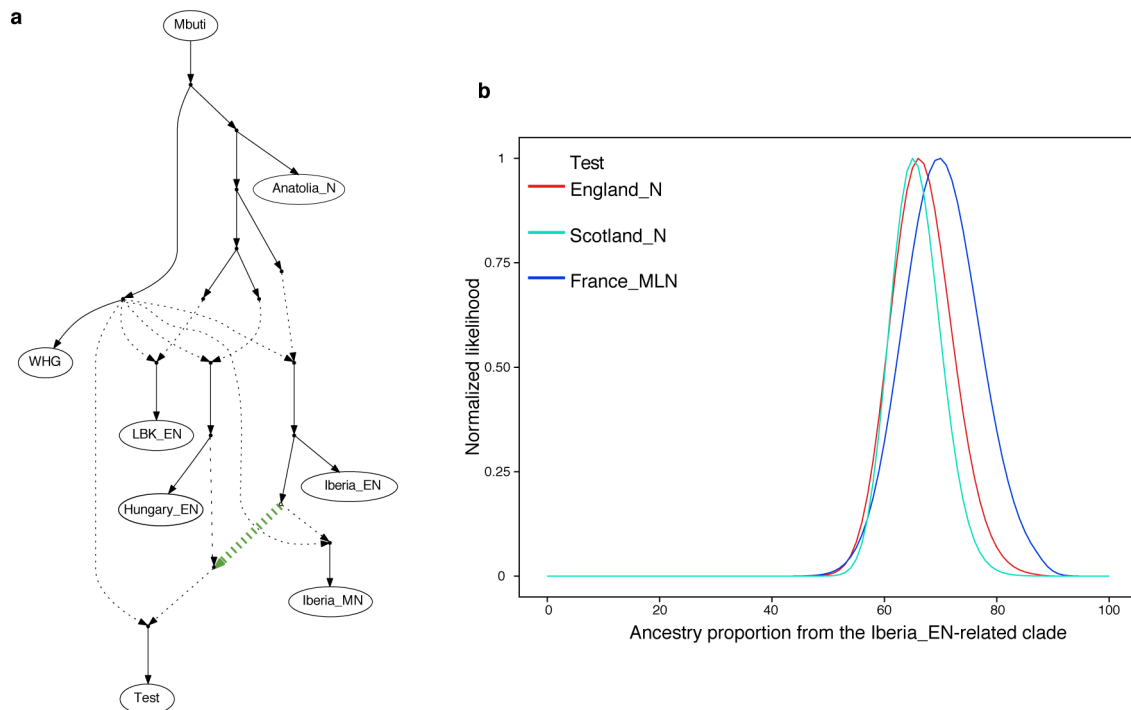
with  $K=8$  showing ancient individuals. WHG, western hunter-gatherers; EHG, eastern hunter-gatherers; SHG, Scandinavian hunter-gatherers; CHG, Caucasus hunter-gatherers; E, Early; M, Middle; L, Late; N, Neolithic; CA, Copper Age; and BA, Bronze Age.



**Extended Data Figure 4 | Hunter-gatherer affinities in Neolithic and Copper Age Europe.** Differential affinity to hunter-gatherer individuals (La Braña1<sup>56</sup> from Spain and KO1<sup>62</sup> from Hungary) in European populations before the emergence of the Beaker complex.

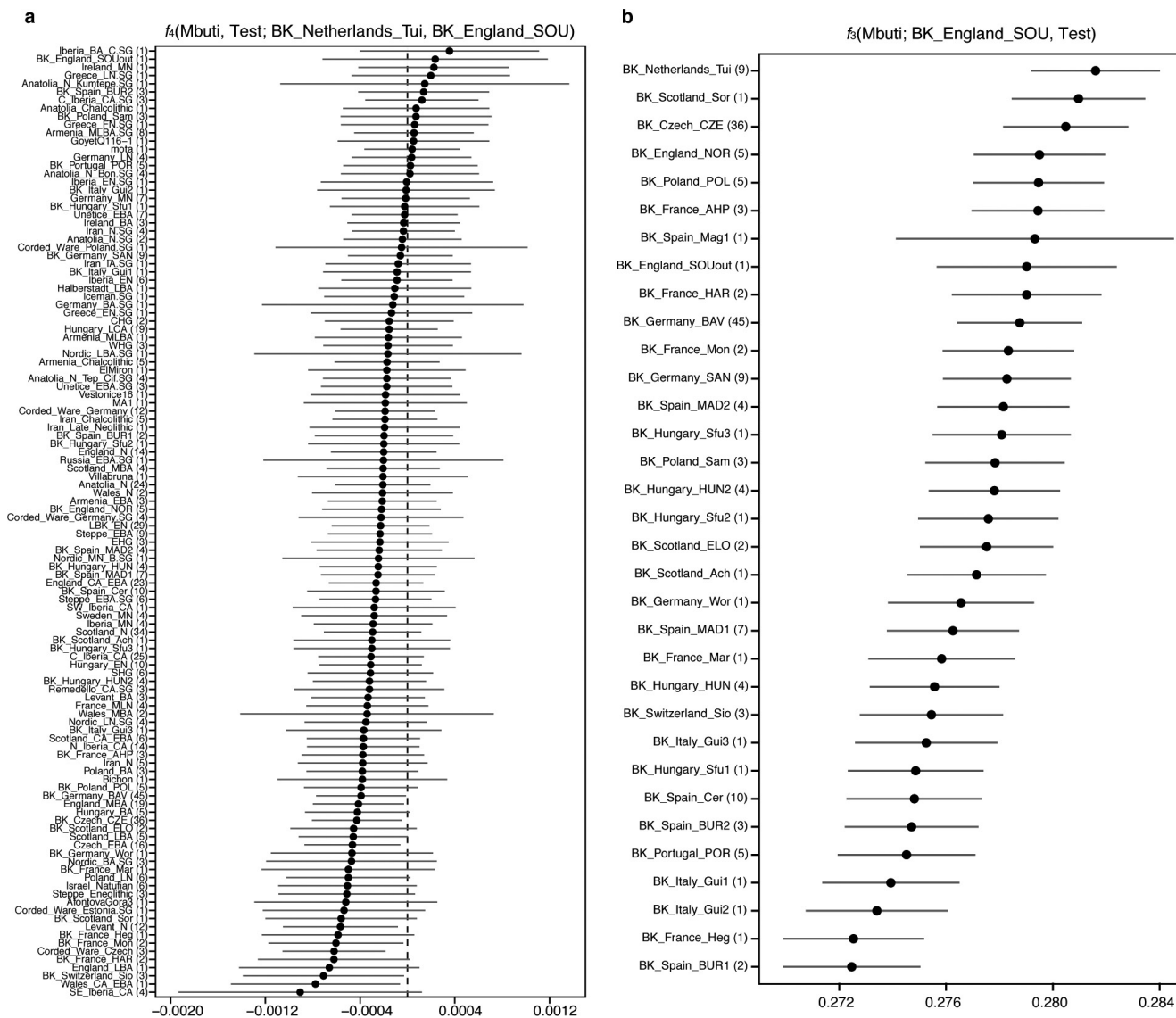
See Supplementary Information section 8 for mixture proportions and standard errors computed with qpAdm<sup>2</sup>. E, Early; M, Middle; L, Late; N, Neolithic; CA, Copper Age; BA, Bronze Age; N\_Iberia, northern Iberia; and C\_Iberia, central Iberia.



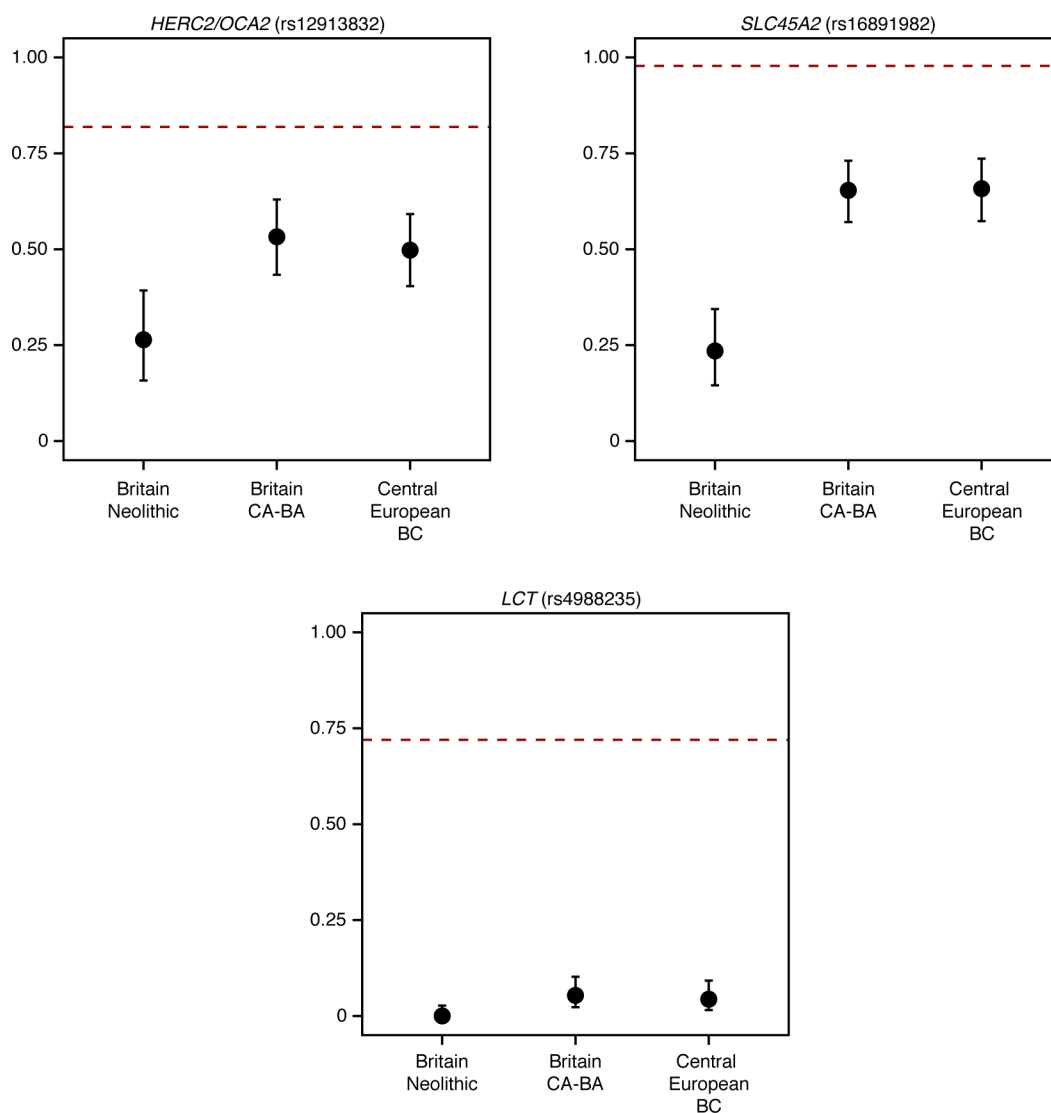


**Extended Data Figure 5 | Modelling the relationships between Neolithic populations.** **a**, Admixture graph fitting a test population as a mixture of sources related to both Iberia\_EN and Hungary\_EN. **b**, Likelihood distribution for models with different proportions of the source related

to Iberia\_EN (green admixture edge in **a**) when the test population is England\_N, Scotland\_N or France\_MLN. E, Early; M, Middle; L, Late; and N, Neolithic.



genetic drift between BK\\_England\\_SOU and other Beaker-complex-associated groups. Error bars represent  $\pm 1$  standard errors. Number of individuals for each group is given in parentheses. BK\\_Netherlands\\_Tui, Beaker-complex-associated individuals from De Tuithoorn (Oostwoud, the Netherlands); BK\\_England\\_SOU, Beaker-complex-associated individuals from southern England. See Supplementary Table 1 for individuals associated with each population label.



**Extended Data Figure 7 | Derived allele frequencies at three SNPs of functional importance.** Error bars represent 1.9-log-likelihood support interval. The red dashed lines show allele frequencies in the 1000 Genomes Project (<http://www.internationalgenome.org/>) 'GBR' population

(present-day people from Great Britain). Sample sizes are 50, 98 and 117 for Britain Neolithic, Britain Copper Age and Bronze Age, and central European Beaker-complex-associated individuals, respectively. BC, Beaker complex; CA, Copper Age; and BA, Bronze Age.



**Extended Data Table 1 | Sites from outside Britain with new genome-wide data reported in this study**

Site	N	Approx. date range (BC)	Country
Brandysek	12	2900–2200	Czech Republic
Kněževes	2	2500–1900	Czech Republic
Lochenice	1	2500–1900	Czech Republic
Lovosice II	1	2500–1900	Czech Republic
Moravská Nová Ves	4	2300–1900	Czech Republic
Prague 5 - Malá Ohrada	1	2500–2200	Czech Republic
Prague 5, Jinonice	14	2200–1700	Czech Republic
Prague 8, Kobylisy, Ke Stírce Street	12	2500–1900	Czech Republic
Radovesice	13	2500–2200	Czech Republic
Velké Přílepy	3	2500–1900	Czech Republic
Clos de Roque, Saint Maximin-la-Sainte-Baume	3	4700–4500	France
Collet Redon, La Couronne-Martigues	1	3500–3100	France
Hégenheim Necropole, Haut-Rhin	1	2800–2500	France
La Fare, Forcalquier	1	2500–2200	France
Marlens, Sur les Barmes, Haute-Savoie	1	2500–2100	France
Mondelange, PAC de la Sente, Moselle	2	2400–1900	France
Rouffach, Haut-Rhin	1	2300–2100	France
Sierentz, Les Villas d'Aurele, Haut-Rhin	2	2600–2300	France
Villard, Lauzet-Ubaye	2	2200–1900	France
Alburg-Lerchenhaid, Spedition Häring, Bavaria	13	2500–2100	Germany
Augsburg Sportgelände, Augsburg, Bavaria	6	2500–2000	Germany
Hugo-Eckener-Straße, Augsburg, Bavaria	3	2500–2000	Germany
Irlbach, County of Straubing-Bogen, Bavaria	17	2500–2000	Germany
Künzing-Bruck, Lkr. Deggendorf, Bavaria	3	2500–2000	Germany
Landau an der Isar, Bavaria	5	2500–2000	Germany
Manching-Oberstimm, Bavaria	2	2500–2000	Germany
Osterhofen-Altenmarkt, Bavaria	4	2600–2000	Germany
Unterer Talweg 58-62, Augsburg, Bavaria	2	2500–2200	Germany
Unterer Talweg 85, Augsburg, Bavaria	1	2400–2100	Germany
Weichering, Bavaria	4	2500–2000	Germany
Worms-Herrnsheim, Rhineland-Palatinate	1	2500–2000	Germany
Budakalász, Csajerszke (M0 Site 12)	2	2600–2200	Hungary
Budapest-Békásmegyer	3	2500–2100	Hungary
Mezőcsát-Hörsöngös	4	3400–3000	Hungary
Szigetszentmiklós-Üdülősor	4	2500–2200	Hungary
Szigetszentmiklós,Felső Űrge-hegyi dűlő	6	2500–2200	Hungary
Pergole 2, Partanna, Sicily	3	2500–1900	Italy
Via Guidorossi, Parma, Emilia Romagna	3	2200–1900	Italy
Dzielnica	1	2300–2000	Poland
Iwiny	1	2300–2000	Poland
Jordanów Śląski	1	2300–2200	Poland
Kornice	4	2500–2100	Poland
Racibórz-Stara Wieś	1	2300–2000	Poland
Samborzec	3	2500–2100	Poland
Strachów	1	2000–1800	Poland
Żerniki Wielkie	1	2300–2100	Poland
Bolores, Estremadura	1	2800–2600	Portugal
Cova da Moura, Torres Vedras	1	2300–2100	Portugal
Galeria da Cisterna, Almonda	2	2500–2200	Portugal
Verdelha dos Ruivos, District of Lisbon	3	2700–2300	Portugal
Arroyal I, Burgos	5	2600–2200	Spain
Camino de las Yeseras, Madrid	14	2800–1700	Spain
Camino del Molino, Caravaca, Murcia	4	2900–2100	Spain
Humanejos, Madrid	11	2900–2000	Spain
La Magdalena, Madrid	3	2500–2000	Spain
Paris Street, Cerdanyola, Barcelona	10	2900–2300	Spain
Virgatal, Tablada de Rudrón, Burgos	1	2300–2000	Spain
Sion-Petit-Chasseur, Dolmen XI	3	2500–2000	Switzerland
De Tuithoorn, Oostwoud, Noord-Holland	11	2600–1600	The Netherlands

Extended Data Table 2 | Sites from Britain with new genome-wide data reported in this study

Site	N	Approx. date range (BC)	Country
Abingdon Spring Road cemetery, Oxfordshire, England	1	2500–2200	Great Britain
Amesbury Down, Wiltshire, England	13	2500–1300	Great Britain
Banbury Lane, Northamptonshire, England	3	3400–3100	Great Britain
Barrow Hills, Radley, Oxfordshire, England	1	2300–1800	Great Britain
Barton Stacey, Hampshire, England	1	2200–2000	Great Britain
Baston and Langtoft, South Lincolnshire, England	2	1700–1600	Great Britain
Biddenham Loop, Bedfordshire, England	9	1600–1300	Great Britain
Boscombe Airfield, Wiltshire, England	1	1800–1600	Great Britain
Canada Farm, Sixpenny Handley, Dorset, England	2	2500–2300	Great Britain
Carsington Pasture Cave, Derbyshire, England	2	3700–2000	Great Britain
Central Flying School, Upavon, Wiltshire, England	1	2500–1800	Great Britain
Cissbury Flint Mine, Worthing, West Sussex, England	1	3600–3400	Great Britain
Clay Farm, Cambridgeshire, England	2	1400–1300	Great Britain
Dairy Farm, Willington, England	1	2300–1900	Great Britain
Ditchling Road, Brighton, Sussex, England	1	2500–1900	Great Britain
Eton Rowing Course, Buckinghamshire, England	2	3600–2900	Great Britain
Flying School, Netheravon, Wiltshire, England	2	2500–1800	Great Britain
Fussell's Lodge, Salisbury, Wiltshire, England	2	3800–3600	Great Britain
Lesser Kelco Cave, Giggleswick Scar, North Yorkshire, England	1	3700–3500	Great Britain
Hasting Hill, Sunderland, Tyne and Wear, England	2	2500–1800	Great Britain
Hexham Golf Course, Northumberland, England	1	2000–1800	Great Britain
Low Hauxley, Northumberland, England	2	2100–1600	Great Britain
Melton Quarry, East Riding of Yorkshire, England	1	1900–1700	Great Britain
Neale's Cave, Paington, Devon, England	1	2000–1600	Great Britain
Nr. Abington, Figheldean, England	1	2500–1800	Great Britain
Nr. Millbarrow, Wiltshire, England	1	3600–3400	Great Britain
Over Narrows, Needingworth Quarry, England	5	2200–1300	Great Britain
Porton Down, Wiltshire, England	2	2500–1900	Great Britain
Raven Scar Cave, Ingleton, North Yorkshire, England	1	1100–900	Great Britain
Reaverhill, Barrasford, Northumberland, England	1	2100–2000	Great Britain
River Thames, Mortlake/Syon Reach, London, England	2	2500–1700	Great Britain
Staxton Beacon, Staxton, England	1	2400–1600	Great Britain
Summerhill, Blaydon, Tyne and Wear, England	1	1900–1700	Great Britain
East Kent Access (Phase II), Thanet, Kent, England	4	2100–1700	Great Britain
Totty Pot, Cheddar, Somerset, England	1	2800–2500	Great Britain
Trumpington Meadows, Cambridge, England	2	2200–2000	Great Britain
Turners Yard, Fordham, Cambridgeshire, England	1	1700–1500	Great Britain
Upper Swell, Chipping Norton, Gloucestershire, England	1	4000–3300	Great Britain
Waterhall Farm, Chippenham, Cambridgeshire, England	1	2000–1700	Great Britain
West Deeping, Lincolnshire, England	1	2300–2000	Great Britain
Whitehawk, Brighton, Sussex, England	1	3700–3400	Great Britain
Wick Barrow, Stogursey, Somerset, England	1	2400–2000	Great Britain
Wilsford Down, Wilsford-cum-Lake, Wiltshire, England	2	2400–2000	Great Britain
Windmill Fields, Stockton-on-Tees, North Yorkshire, England	4	2300–2000	Great Britain
Yarnton, Oxfordshire, England	4	2500–1900	Great Britain
Aberdour Road, Dunfermline, Fife, Scotland	1	2000–1800	Great Britain
Achavanich, Wick, Highland, Scotland	1	2500–2100	Great Britain
Boatbridge Quarry, Thankerton, Scotland	1	2400–2100	Great Britain
Clachaig, Arran, North Ayrshire, Scotland	1	3500–3400	Great Britain
Covesea Cave 2, Moray, Scotland	3	2100–800	Great Britain
Covesea Caves, Moray, Scotland	2	1000–800	Great Britain
Distillery Cave, Oban, Argyll and Bute, Scotland	3	3800–3400	Great Britain
Doune, Perth and Kinross, Scotland	1	1800–1600	Great Britain
Dryburn Bridge, East Lothian, Scotland	2	2300–1900	Great Britain
Eweford Cottages, East Lothian, Scotland	1	2100–1900	Great Britain
Holm of Papa Westray North, Orkney, Scotland	4	3500–3100	Great Britain
Isbister, Orkney, Scotland	10	3300–2300	Great Britain
Leith, Merrilees Close, City of Edinburgh, Scotland	2	1600–1500	Great Britain
Longniddry, Evergreen House, Coast Road, East Lothian, Scotland	3	1500–1300	Great Britain
Longniddry, Grainfoot, East Lothian, Scotland	1	1300–1000	Great Britain
Macarthur Cave, Oban, Argyll and Bute, Scotland	1	4000–3800	Great Britain
Pabay Mor, Lewis, Western Isles, Scotland	1	1400–1300	Great Britain
Point of Cott, Orkney, Scotland	2	3700–3100	Great Britain
Quoyness, Orkney, Scotland	1	3100–2900	Great Britain
Raschoille Cave, Oban, Argyll and Bute, Scotland	9	4000–2900	Great Britain
Sorisdale, Coll, Argyll and Bute, Scotland	1	2500–2100	Great Britain
Stenchme, Lop Ness, Orkney, Scotland	1	2000–1500	Great Britain
Thurston Mains, Innerwick, East Lothian, Scotland	1	2300–2000	Great Britain
Tulach an t'Sionnach, Highland, Scotland	1	3700–3500	Great Britain
Tulloch of Assery A, Highland, Scotland	1	3700–3400	Great Britain
Tulloch of Assery B, Highland, Scotland	1	3800–3600	Great Britain
Unstan, Orkney, Scotland	1	3400–3100	Great Britain
Culver Hole Cave, Port Eynon, West Glamorgan, Wales	1	1600–800	Great Britain
Great Orme Mines, Llandudno, North Wales	1	1700–1600	Great Britain
North Face Cave, Llandudno, North Wales	1	1400–1200	Great Britain
Rhos Ddigre, Llanarmon-yn-Iâl, Denbighshire, Wales	1	3100–2900	Great Britain
Tinkinswood, Cardiff, Glamorgan, Wales	1	3800–3600	Great Britain



Extended Data Table 3 | 111 newly reported radiocarbon dates

Sample	Date		Location	Country
I5024	2278–2032 calBC	(3740±35 BP, Poz-84460)	Kněževy	Czech Republic
I4946	2296–2146 calBC	(3805±20 BP, PSUAMS-2801)	Prague 5, Jinonice, Butovická Street	Czech Republic
I4895	2273–2047 calBC	(3750±20 BP, PSUAMS-2852)	Prague 5, Jinonice, Butovická Street	Czech Republic
I4896	2288–2142 calBC	(3785±20 BP, PSUAMS-2853)	Prague 5, Jinonice, Butovická Street	Czech Republic
I4884	1882–1745 calBC	(3480±20 BP, PSUAMS-2842)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4885	2289–2143 calBC	(3790±20 BP, PSUAMS-2843)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4886	2205–2042 calBC	(3740±20 BP, PSUAMS-2844)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4887	2201–2039 calBC	(3730±20 BP, PSUAMS-2845)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4888	2190–2029 calBC	(3700±20 BP, PSUAMS-2846)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4889	2281–2062 calBC	(3765±20 BP, PSUAMS-2847)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4891	2281–2062 calBC	(3765±20 BP, PSUAMS-2848)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4892	1881–1701 calBC	(3475±20 BP, PSUAMS-2849)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4893	4449–4348 calBC	(5550±20 BP, PSUAMS-2850)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4894	4488–4368 calBC	(5610±20 BP, PSUAMS-2851)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4945	2291–2144 calBC	(3795±20 BP, PSUAMS-2854)	Prague 8, Kobylisy, Ke Stírci Street	Czech Republic
I4305	4825–4616 calBC	(5860±35 BP, PSUAMS-2225)	Clos de Roque, Saint Maximin-la-Sainte-Baume	France
I4304	4787–4589 calBC	(5830±35 BP, PSUAMS-2226)	Clos de Roque, Saint Maximin-la-Sainte-Baume	France
I4303	4778–4586 calBC	(5820±30 BP, PSUAMS-2260)	Clos de Roque, Saint Maximin-la-Sainte-Baume	France
I1392	2833–2475 calBC	(4047±29 BP, MAMS-25935)	Hörsheim Necropole, Haut-Rhin	France
I3875	2133–1948 calBC	(3655±25 BP, PSUAMS-1834)	Villard, Lauzet-Ubaye	France
I3874	2200–2035 calBC	(3725±25 BP, PSUAMS-1835)	Villard, Lauzet-Ubaye	France
I3593	2397–2145 calBC	(3817±26 BP, BRAMS-1215)	Alburg-Lerchenhaid, Spedition Häring, Stkr. Straubing, Bavaria	Germany
I3590	2335–2140 calBC	(3802±26 BP, BRAMS-1217)	Alburg-Lerchenhaid, Spedition Häring, Stkr. Straubing, Bavaria	Germany
I3592	2457–2203 calBC	(3844±33 BP, BRAMS-1218)	Alburg-Lerchenhaid, Spedition Häring, Stkr. Straubing, Bavaria	Germany
I5017	2460–2206 calBC	(3855±35 BP, Poz-84458)	Augsburg Sportgelände, Augsburg, Bavaria	Germany
I4250	2433–2149 calBC	(3825±26 BP, BRAMS-1219)	Irlbach, County of Straubing-Bogen, Bavaria	Germany
I5021	2571–2341 calBC	(3955±35 BP, Poz-84553)	Osterhofen-Altenmarkt, Bavaria	Germany
E09537_d	2471–2298 calBC	(3909±29 BP, MAMS-29074)	Unterer Talweg 58-62, Augsburg, Bavaria	Germany
E09538	2468–2210 calBC	(3870±30 BP, MAMS-29075)	Unterer Talweg 58-62, Augsburg, Bavaria	Germany
I5385	2455–2147 calBC	(3827±33 BP, SUERC-71005)	Achavanich, Wick, Highland, Scotland	Great Britain
I2457	2199–2030 calBC	(3717±28 BP, SUERC-69975)	Amesbury Down, Wiltshire, England	Great Britain
I2416	2455–2151 calBC	(3830±30 BP, Beta-432804)	Amesbury Down, Wiltshire, England	Great Britain
I2596	2273–2034 calBC	(3739±30 BP, NZA-32484)	Amesbury Down, Wiltshire, England	Great Britain
I2566	2204–2035 calBC	(3734±25 BP, NZA-32490)	Amesbury Down, Wiltshire, England	Great Britain
I2598	2135–1953 calBC	(3664±30 BP, NZA-32494)	Amesbury Down, Wiltshire, England	Great Britain
I2418	2455–2200 calBC	(3836±25 BP, NZA-32788)	Amesbury Down, Wiltshire, England	Great Britain
I2565	2457–2147 calBC	(3829±38 BP, OxA-13562)	Amesbury Down, Wiltshire, England	Great Britain
I2457	2467–2290 calBC	(3890±30 BP, SUERC-36210)	Amesbury Down, Wiltshire, England	Great Britain
I2460	2461–2187 calBC	(3875±27 BP, SUERC-5304)	Amesbury Down, Wiltshire, England	Great Britain
I2459	2455–2150 calBC	(3829±30 BP, SUERC-54823)	Amesbury Down, Wiltshire, England	Great Britain
I5373	2194–1980 calBC	(3694±25 BP, BRAMS-1230)	Carsington Pasture Cave, Brassington, Derbyshire, England	Great Britain
I2988	3516–3361 calBC	(4645±29 BP, SUERC-68711)	Clachaig, Arran, North Ayrshire, Scotland	Great Britain
I2860	969–815 calBC	(2738±29 BP, SUERC-68715)	Covesea Cave 2, Moray, Scotland	Great Britain
I2861	976–828 calBC	(2757±29 BP, SUERC-68716)	Covesea Cave 2, Moray, Scotland	Great Britain
I3132	2118–1887 calBC	(3614±33 BP, SUERC-69070)	Covesea Cave 2, Moray, Scotland	Great Britain
I3130	977–829 calBC	(2758±29 BP, SUERC-68713)	Covesea Caves, Moray, Scotland	Great Britain
I2859	910–809 calBC	(2714±29 BP, SUERC-68714)	Covesea Caves, Moray, Scotland	Great Britain
I2452	2198–1980 calBC	(3700±30 BP, Beta-444979)	Dairy Farm, Willington, England	Great Britain
I2452	2276–2029 calBC	(3730±35 BP, Poz-83405)	Dairy Farm, Willington, England	Great Britain
I2659	3761–3643 calBC	(491±27 BP, SUERC-68702)	Distillery Cave, Oban, Argyll and Bute, Scotland	Great Britain
I2660	3513–3352 calBC	(4631±29 BP, SUERC-68703)	Distillery Cave, Oban, Argyll and Bute, Scotland	Great Britain
I2691	3700–3639 calBC	(4881±25 BP, SUERC-68704)	Distillery Cave, Oban, Argyll and Bute, Scotland	Great Britain
I6774	2287–2044 calBC	(3760±30 BP, SUERC-74755)	Ditchling Road, Brighton, Sussex, England	Great Britain
I2605	3631–3372 calBC	(4710±35 BP, Poz-83483)	Eton Rowing Course, Buckinghamshire, England	Great Britain
I1775	1730–1532 calBC	(3344±27 BP, OxA-14308)	Great Orme, Llandudno, North Wales	Great Britain
I2574	1414–1227 calBC	(3065±36 BP, SUERC-62072)	Great Orme, Llandudno, North Wales	Great Britain
I2612	2464–2208 calBC	(3865±35 BP, Poz-83492)	Hasting Hill, Sunderland, Tyne and Wear, England	Great Britain
I2609	2022–1771 calBC	(3560±40 BP, Poz-83423)	Hexham Golf Course, Northumberland, England	Great Britain
I2636	3519–3381 calBC	(4651±33 BP, SUERC-68640)	Holm of Papa Westray North, Orkney, Scotland	Great Britain
I2637	3629–3370 calBC	(4697±33 BP, SUERC-68641)	Holm of Papa Westray North, Orkney, Scotland	Great Britain
I2650	3638–3380 calBC	(4754±36 BP, SUERC-68642)	Holm of Papa Westray North, Orkney, Scotland	Great Britain
I2651	3360–3098 calBC	(4525±36 BP, SUERC-68643)	Holm of Papa Westray North, Orkney, Scotland	Great Britain
I2630	2580–2463 calBC	(3999±32 BP, SUERC-68632)	Isbister, Orkney, Scotland	Great Britain
I2932	2570–2347 calBC	(3962±29 BP, SUERC-68721)	Isbister, Orkney, Scotland	Great Britain
I2933	3010–2885 calBC	(4309±29 BP, SUERC-68722)	Isbister, Orkney, Scotland	Great Britain
I2935	3335–3011 calBC	(4451±29 BP, SUERC-68723)	Isbister, Orkney, Scotland	Great Britain
I3085	3338–3026 calBC	(4471±29 BP, SUERC-68724)	Isbister, Orkney, Scotland	Great Britain
I2978	3335–3023 calBC	(4464±29 BP, SUERC-68725)	Isbister, Orkney, Scotland	Great Britain
I2979	3294–2941 calBC	(4447±29 BP, SUERC-68726)	Isbister, Orkney, Scotland	Great Britain
I2934	3338–3022 calBC	(4466±33 BP, SUERC-69071)	Isbister, Orkney, Scotland	Great Britain
I2977	3008–2763 calBC	(4275±33 BP, SUERC-69072)	Isbister, Orkney, Scotland	Great Britain
I2657	3951–3780 calBC	(5052±30 BP, SUERC-68701)	Macarthur Cave, Oban, Argyll and Bute, Scotland	Great Britain
I5441	1938–1744 calBC	(351±23 BP, OxA-16522)	Neale's Cave, Paington, Devon, England	Great Britain
I4949	3629–3376 calBC	(4715±20 BP, PSUAMS-2513)	Nr. Millbarrow, Winterbourne Monkton, Wiltshire, England	Great Britain
I2980	3360–3101 calBC	(4530±33 BP, SUERC-69073)	Point of Cott, Orkney, Scotland	Great Britain
I2796	3705–3535 calBC	(4856±33 BP, SUERC-69074)	Point of Cott, Orkney, Scotland	Great Britain
I2631	3097–2906 calBC	(4384±36 BP, SUERC-68633)	Quoyness, Orkney, Scotland	Great Britain
I3135	3640–3383 calBC	(4770±30 BP, PSUAMS-2068)	Raschoille Cave, Oban, Argyll and Bute, Scotland	Great Britain
I3136	3520–3365 calBC	(4665±30 BP, PSUAMS-2069)	Raschoille Cave, Oban, Argyll and Bute, Scotland	Great Britain
I3133	3631–3377 calBC	(4725±20 BP, PSUAMS-2154)	Raschoille Cave, Oban, Argyll and Bute, Scotland	Great Britain
I3134	3633–3377 calBC	(4730±25 BP, PSUAMS-2155)	Raschoille Cave, Oban, Argyll and Bute, Scotland	Great Britain
I3138	3263–2923 calBC	(4415±25 BP, PSUAMS-2156)	Raschoille Cave, Oban, Argyll and Bute, Scotland	Great Britain
I2610	1935–1745 calBC	(351±35 BP, Poz-83498)	Summerhill, Blaydon, Tyne and Wear, England	Great Britain
I2634	3703–3534 calBC	(4851±34 BP, SUERC-68638)	Tulach an t'Sionnach, Highland, Scotland	Great Britain
I2635	3652–3389 calBC	(4796±37 BP, SUERC-68639)	Tulloch of Assery A, Highland, Scotland	Great Britain
I2633	3765–3641 calBC	(4911±32 BP, SUERC-68634)	Tulloch of Assery B, Highland, Scotland	Great Britain
I2453	2288–2040 calBC	(3760±35 BP, Poz-83404)	West Deeping, Lincolnshire, England	Great Britain
I2445	2136–1929 calBC	(3650±35 BP, Poz-83407)	Yarnton, Oxfordshire, England	Great Britain
I2447	2115–1910 calBC	(3625±25 BP, PSUAMS-2336)	Yarnton, Oxfordshire, England	Great Britain
I2786	2458–2205 calBC	(3850±35 BP, Poz-83639)	Szigetszentmiklós-Felső-Urge hegyi dűlő	Hungary
I2787	2457–2201 calBC	(3840±35 BP, Poz-83640)	Szigetszentmiklós-Felső-Urge hegyi dűlő	Hungary
I2741	2457–2153 calBC	(3835±35 BP, Poz-83641)	Szigetszentmiklós-Felső-Urge hegyi dűlő	Hungary
I6531	2286–2038 calBC	(3755±35 BP, Poz-86947)	Dzielnica	Poland
I6579	2335–2046 calBC	(3780±35 BP, Poz-75954)	Iwiny	Poland
I6534	2456–2149 calBC	(3830±35 BP, Poz-75936)	Kornice	Poland
I6582	2343–2057 calBC	(3790±35 BP, Poz-75951)	Kornice	Poland
I4251	2431–2150 calBC	(3825±25 BP, PSUAMS-2321)	Samborzec 1	Poland
I4252	2285–2138 calBC	(3780±20 BP, PSUAMS-2338)	Samborzec 1	Poland
I4253	2456–2207 calBC	(3865±20 BP, PSUAMS-2339)	Samborzec 1	Poland
I6538	2008–1765 calBC	(3545±35 BP, Poz-86950)	Strachów	Poland
I6583	2289–2050 calBC	(3770±30 BP, Poz-65207)	Zerniki Wielkie	Poland
I4229	2288–2134 calBC	(3775±25 BP, PSUAMS-1750)	Cova da Moura	Portugal
I0462	2566–2345 calBC	(3950±26 BP, MAMS-25936)	Arroyal I, Burgos	Spain
I4247	2464–2210 calBC	(3870±30 BP, PSUAMS-2120)	Camino de las Yeseras, Madrid	Spain
I4245	2460–2291 calBC	(3875±20 BP, PSUAMS-2320)	Camino de las Yeseras, Madrid	Spain
I0257	2572–2348 calBC	(3965±29 BP, MAMS-25937)	Paris Street, Cerdanyola, Barcelona	Spain
I0825	2474–2298 calBC	(3915±29 BP, MAMS-25939)	Paris Street, Cerdanyola, Barcelona	Spain
I0826	2834–2482 calBC	(4051±28 BP, MAMS-25940)	Paris Street, Cerdanyola, Barcelona	Spain
I4068	2131–1951 calBC	(3655±20 BP, PSUAMS-2318)	De Tuithoorn, Oostwoud, Noord-Holland	The Netherlands
I4076	1882–1750 calBC	(3490±20 BP, PSUAMS-2319)	De Tuithoorn, Oostwoud, Noord-Holland	The Netherlands
I4075	2118–1937 calBC	(3635±20 BP, PSUAMS-2337)	De Tuithoorn, Oostwoud, Noord-Holland	The Netherlands



# The genomic history of southeastern Europe

A list of authors and affiliations appears at the end of the paper.

**Farming was first introduced to Europe in the mid-seventh millennium BC, and was associated with migrants from Anatolia who settled in the southeast before spreading throughout Europe. Here, to understand the dynamics of this process, we analysed genome-wide ancient DNA data from 225 individuals who lived in southeastern Europe and surrounding regions between 12000 and 500 BC. We document a west-east cline of ancestry in indigenous hunter-gatherers and, in eastern Europe, the early stages in the formation of Bronze Age steppe ancestry. We show that the first farmers of northern and western Europe dispersed through southeastern Europe with limited hunter-gatherer admixture, but that some early groups in the southeast mixed extensively with hunter-gatherers without the sex-biased admixture that prevailed later in the north and west. We also show that southeastern Europe continued to be a nexus between east and west after the arrival of farmers, with intermittent genetic contact with steppe populations occurring up to 2,000 years earlier than the migrations from the steppe that ultimately replaced much of the population of northern Europe.**

Southeastern Europe was the beachhead in the spread of agriculture from its source in the Fertile Crescent of southwestern Asia. After the first appearance of agriculture in Europe in the mid-seventh millennium BC<sup>1,2</sup>, farming spread westward along a Mediterranean route, and northwestward via a Danubian route, and was established in Iberia and central Europe by 5600 BC<sup>3,4</sup>. Previous studies have shown that the spread of farming across Europe was accompanied by a massive movement of people<sup>5–8</sup> who were closely related to the farmers of northwestern Anatolia<sup>9–11</sup>, but nearly all the ancient DNA on which these studies are based derives from first farmers in central and western Europe, with only three individuals reported from the southeast<sup>9</sup>. In the two millennia after the establishment of agriculture in the Balkan Peninsula a series of complex societies formed, which culminated in sites such as the mid-fifth millennium BC necropolis at Varna. The Varna necropolis has some of the earliest evidence for extreme inequality in wealth; one individual there (grave 43), from whom we extracted DNA, was buried with more gold than is known from all other Neolithic and Copper Age burials, combined. By the end of the sixth millennium BC, agriculture had reached eastern Europe, where it is associated with the Cucuteni–Trypillian complex in the area of present-day Moldova, Romania and Ukraine; this complex was characterized by ‘mega-sites’ that housed hundreds, or perhaps even thousands, of people<sup>12</sup>. After around 4000 BC, these settlements were largely abandoned, and archaeological evidence documents cultural contacts with peoples of the Eurasian steppe<sup>13</sup>. However, the population movements that accompanied these events have previously been unknown, owing to the lack of ancient DNA evidence.

## Results

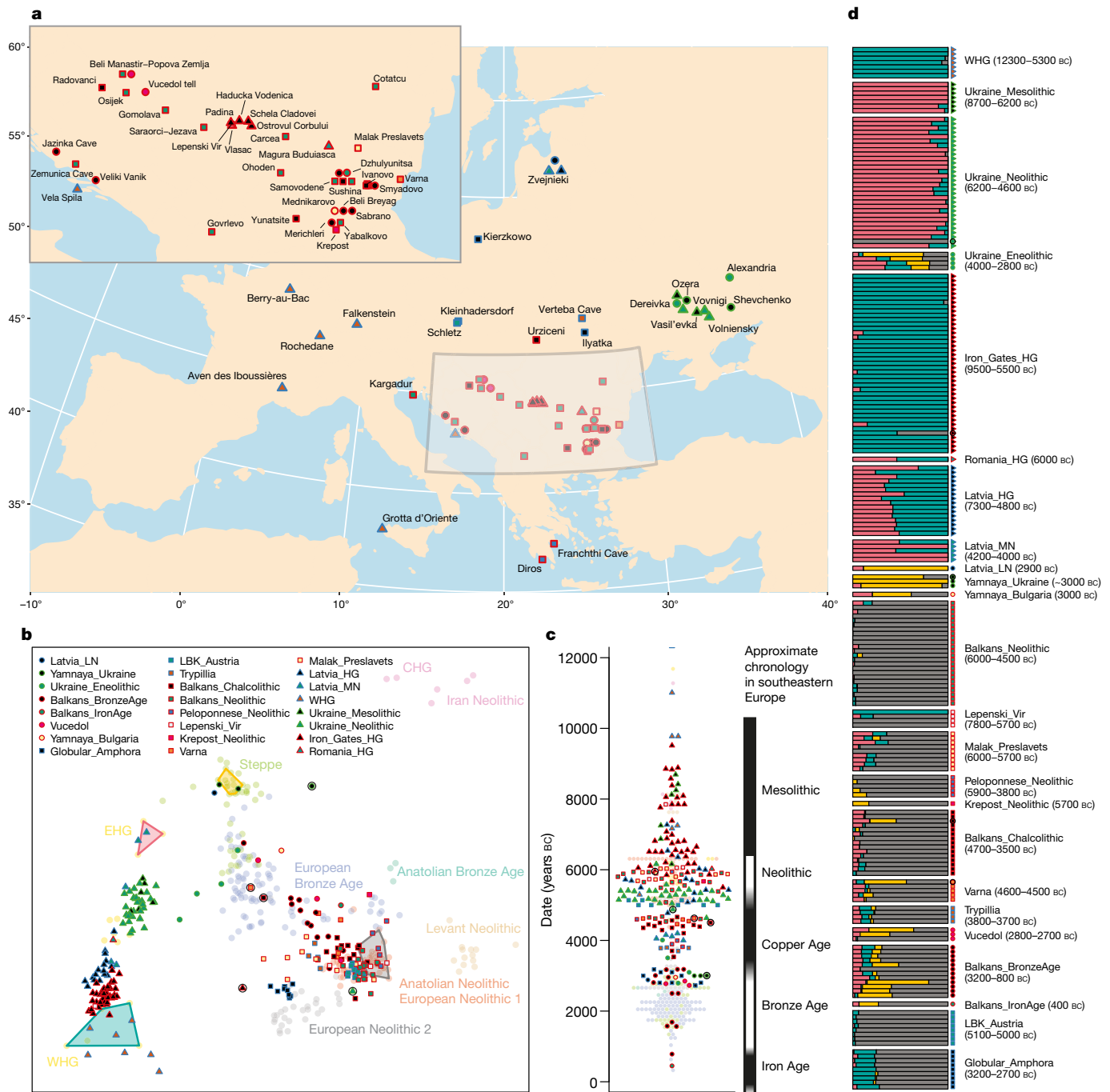
We generated genome-wide data from 225 ancient humans dated to 12000–500 BC, from the Balkan Peninsula, the Carpathian Basin, the North Pontic steppe and neighbouring regions (Fig. 1, Supplementary Table 1, Supplementary Note 1); 215 of these individuals are reported here for the first time. We extracted DNA from skeletal remains in dedicated clean rooms, built DNA libraries and enriched for DNA fragments overlapping 1.24 million single nucleotide polymorphisms (SNPs), and then sequenced the product and restricted to libraries with evidence of authentic ancient DNA<sup>7,10,14</sup>. We filtered

out individuals with fewer than 15,000 SNPs covered by at least one sequence, or those that had unexpected ancestry for their archaeological context and were not directly dated. We report, but do not analyse, nine individuals that were first-degree relatives of others in the dataset, resulting in an analysis dataset of 216 individuals. We analysed these data together with data from 274 previously reported ancient individuals<sup>9–11,15–27</sup>, 777 present-day individuals genotyped on the Illumina ‘Human Origins’ array<sup>23</sup> and 300 high-coverage genomes from the Simons Genome Diversity Project<sup>28</sup>. We used principal component analysis (Fig. 1b, Extended Data Fig. 1), supervised and unsupervised ADMIXTURE<sup>29</sup> (Fig. 1d, Extended Data Figs 2, 3), *D* statistics, qpAdm and qpGraph<sup>30</sup>, along with archaeological and chronological information (including 137 newly reported accelerator mass spectrometry (AMS) <sup>14</sup>C dates) to cluster the individuals into populations and investigate the relationships among them.

We described the individuals in our dataset in terms of their genetic relatedness to a hypothesized set of ancestral populations, which we refer to as their genetic ancestry. It has previously been shown that the great majority of European ancestry derives from three distinct sources<sup>23</sup>: first, ‘hunter-gatherer-related’ ancestry that is more closely related to Mesolithic hunter-gatherers from Europe than to any other population, and that can be further subdivided into ‘eastern’ (EHG) and ‘western’ (WHG) hunter-gatherer-related ancestry<sup>7</sup>; second, ‘northwestern-Anatolian-Neolithic-related’ ancestry related to the Neolithic farmers of northwest Anatolia and tightly linked to the appearance of agriculture<sup>9,10</sup>; and third, ‘steppe-related’ ancestry that appears in western Europe during the Late Neolithic-to-Bronze Age transition, and which is ultimately derived from a population related to Yamnaya steppe pastoralists<sup>7,15</sup>. Steppe-related ancestry itself can be modelled as a mixture of EHG-related ancestry and ancestry related to Upper Palaeolithic hunter-gatherers of the Caucasus (CHG) and the first farmers of northern Iran<sup>19,21,22</sup>.

## Hunter-gatherer substructure and transitions

Of the 215 new individuals that we report, 105 from Palaeolithic, Mesolithic and eastern European Neolithic contexts (unlike in western Europe, the eastern European Neolithic refers to the presence



**Figure 1 | Geographic and genetic structure of 216 analysed individuals.** **a**, Locations of newly reported individuals. **b**, Ancient individuals projected onto principal components defined by 777 present-day west Eurasians (shown in Extended Data Fig. 1); data include selected published individuals (faded circles, labelled) and newly reported individuals (other symbols, outliers enclosed in black circles). Coloured polygons cover individuals that had cluster memberships fixed at 100% for supervised ADMIXTURE analysis. **c**, Direct or contextual dates

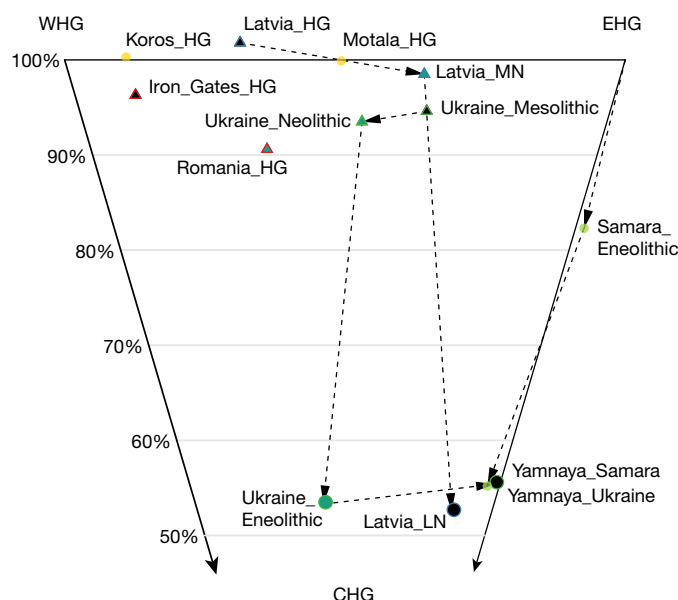
of pottery<sup>31–33</sup>, not necessarily to farming) have almost entirely hunter-gatherer-related ancestry. These individuals form a cline from WHG to EHG that is correlated with geography (Fig. 1b)—although it is neither geographically nor temporally uniform (Fig. 2, Extended Data Fig. 4)—and contains substructure in phenotypically important variants (Supplementary Note 2).

From present-day Ukraine, our study reports genome-wide data from 7 Mesolithic (approximately 9500–6000 bc) and 30 Neolithic

for each sample and approximate chronology of southeastern Europe.

**d**, Supervised ADMIXTURE analysis, modelling each ancient individual (one per row) as a mixture of population clusters constrained to contain northwestern-Anatolian Neolithic (grey), Yamnaya from Samara (yellow), EHG (pink) and WHG (green) populations. Dates in parentheses indicate approximate range of individuals in each population. See Extended Data Fig. 2 for individual sample identification numbers. Map data in **a** from the R package 'maps'.

(approximately 6000–3500 bc) individuals. On the cline from WHG-to-EHG-related ancestry, the Mesolithic individuals fall towards the east, intermediate between EHG and Mesolithic hunter-gatherers from Scandinavia<sup>7</sup> (Fig. 1b). The Neolithic population has a significant difference in ancestry compared to the Mesolithic (Figs 1b, 2), with a shift towards WHG shown by the statistic  $D(\text{Mbuti}, \text{WHG}, \text{Ukraine\_Mesolithic}, \text{Ukraine\_Neolithic})$ ;  $Z = 8.5$  (Supplementary Table 2). Unexpectedly, one Neolithic individual from Dereivka (I3719)



**Figure 2 | Structure and change in hunter-gatherer-related populations.** Inferred ancestry proportions for populations modelled as a mixture of WHG, EHG and CHG (Supplementary Table 3.1.3). Dashed lines show populations from the same geographic region. Percentages indicate proportion of WHG + EHG ancestry. Standard errors range from 1.5 to 8.3% (Supplementary Table 3.1.3).

that we directly dated to 4949–4799 BC has entirely northwestern-Anatolian-Neolithic-related ancestry.

The pastoralist Bronze Age Yamnaya complex originated on the Eurasian steppe and is a plausible source for the dispersal of steppe-related ancestry into central and western Europe from around 2500 BC<sup>13</sup>. All previously reported Yamnaya individuals were from Samara<sup>7</sup> and Kalmykia<sup>15</sup> in southwest Russia, and had entirely steppe-related ancestry. Here, we report three Yamnaya individuals from further west, in Ukraine and Bulgaria, and show that although they all have high levels of steppe-related ancestry, one individual from Ozero in Ukraine and one from Bulgaria (I1917 and Bul4, respectively, both dated to approximately 3000 BC) have northwestern-Anatolian-Neolithic-related admixture, the first evidence of such ancestry in Yamnaya-associated individuals (Fig. 1b, d, Supplementary Table 2). Preceding the Yamnaya complex, four Copper Age individuals (I4110, I5882, I5884 and I6561; labelled as ‘Ukraine\_Eneolithic’) from Dereivka and Alexandria dated to approximately 3600–3400 BC have a mixture of hunter-gatherer-, steppe- and northwestern-Anatolian-Neolithic-related ancestry (Fig. 1d, Supplementary Table 2).

At Zvejnieki in Latvia, using 17 newly reported individuals and additional data for 5 previously reported<sup>34</sup> individuals, we observe a transition in hunter-gatherer-related ancestry that is opposite to that seen in Ukraine. We find that Mesolithic and Early Neolithic individuals (labelled ‘Latvia\_HG’) associated with the Kunda and Narva cultures have ancestry that is intermediate between WHG (approximately 70%) and EHG (approximately 30%), consistent with previous reports<sup>34–36</sup> (Supplementary Table 3). We also detect a shift in ancestry between Early Neolithic individuals and those associated with the Middle Neolithic Comb Ware complex (labelled ‘Latvia\_MN’), who have more EHG-related ancestry; we estimate that the ancestry of Latvia\_MN individuals comprises 65% EHG-related ancestry, but two of the four individuals appear to be 100% EHG in principal component space (Fig. 1b). The most recent individual, associated with the Final Neolithic Corded Ware complex (I4629, labelled ‘Latvia\_LN’), attests to another ancestry shift, clustering closely with Yamnaya from Samara<sup>7</sup>, Kalmykia<sup>15</sup> and Ukraine (Fig. 2).

We report Upper Palaeolithic and Mesolithic data from southern and western Europe<sup>17</sup>. Sicilian (I2158) and Croatian (I1875) individuals

dating to approximately 12000 and 6100 BC cluster with previously reported WHG (Fig. 1b, d), including individuals from Loschbour<sup>23</sup> (Luxembourg, 6100 BC), Bichon<sup>19</sup> (Switzerland, 11700 BC), and Villabruna<sup>17</sup> (Italy, 12000 BC). These results demonstrate that, for at least six thousand years, WHG populations<sup>23</sup> were widely distributed from the Atlantic seaboard of Europe in the west, to Sicily in the south, and to the Balkan Peninsula in the southeast.

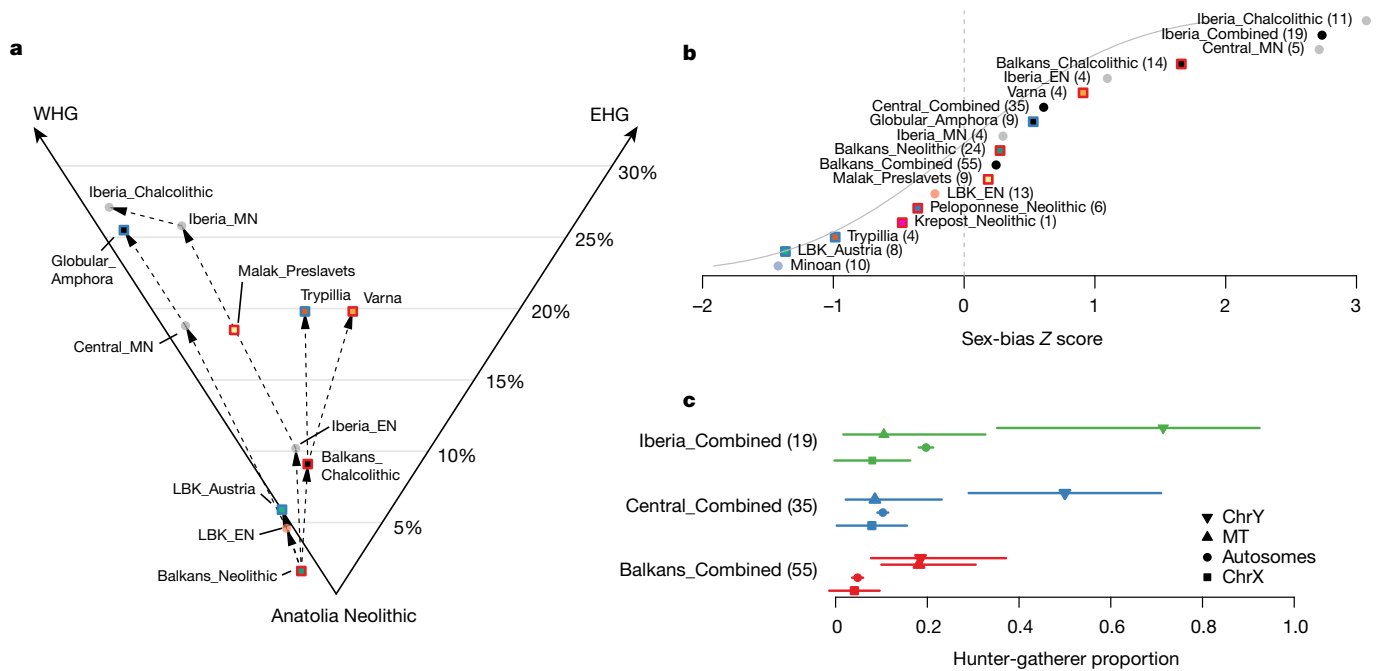
A particularly important hunter-gatherer population that we report is from the Iron Gates region, which straddles the border of present-day Romania and Serbia. This population (labelled ‘Iron\_Gates\_HG’) is represented in our study by 40 individuals from 5 sites. We modelled Iron Gates hunter-gatherers as a mixture of WHG- and EHG-related ancestry (Supplementary Table 3), which showed that they are intermediate between the two (WHG contributing approximately 85%, and EHG approximately 15%, of ancestry). However, this qpAdm model does not fit well ( $P = 0.0003$ , Supplementary Table 3) and the Iron Gates hunter-gatherers show an affinity towards northwestern-Anatolian-Neolithic-, relative to WHG-, ancestry populations (Supplementary Table 2). In addition, Iron Gates hunter-gatherers carry mitochondrial haplogroup K1 (7/40) as well as other subclades of haplogroups U (32/40) and H (1/40), in contrast to WHG, EHG and Scandinavian hunter-gatherers—almost all of whom carry haplogroups U5 or U2. One interpretation is that the Iron Gates hunter-gatherers have ancestry that is not present in either WHG or EHG. Possible scenarios include genetic contact between the ancestors of the Iron Gates population and a northwestern-Anatolian-Neolithic-related population, or that the Iron Gates population is related to the source population from which the WHG split during a re-expansion into Europe from the southeast after the Last Glacial Maximum<sup>17,37</sup>.

A notable finding from the Iron Gates concerns the four individuals from the site of Lepenski Vir, two of whom (I4665 and I5405, 6200–5600 BC), have entirely northwestern-Anatolian-Neolithic-related ancestry. Strontium and nitrogen isotope data<sup>38</sup> indicate that both these individuals were migrants from outside the Iron Gates region and ate a primarily terrestrial diet (Supplementary Information section 1). A third individual (I4666, 6070 BC) has a mixture of northwestern-Anatolian-Neolithic-related and hunter-gatherer-related ancestry and consumed aquatic foods, and a fourth and probably earlier individual (I5407) had entirely hunter-gatherer-related ancestry (Fig. 1d, Supplementary Information section 1). We also identify one individual from Padina (I5232), dated to 5950 BC, who had a mixture of northwestern-Anatolian-Neolithic-related and hunter-gatherer-related ancestry. These results provide genetic confirmation that the Iron Gates was a region of interaction between groups distinct in both ancestry and subsistence strategy.

### Population transformations in the first farmers

Neolithic populations from present-day Bulgaria, Croatia, the former Yugoslav Republic of Macedonia, Serbia and Romania cluster closely with the northwestern-Anatolian-Neolithic individuals (Fig. 1), consistent with archaeological evidence<sup>39</sup>. Modelling Balkan Neolithic populations as a mixture of northwestern-Anatolian-Neolithic and WHG, we estimate that 98% (95% confidence interval; 97–100%) of their ancestry is northwestern-Anatolian-Neolithic-related. A notable exception is evident in eight out of nine individuals from Malak Preslavets in present-day Bulgaria<sup>40</sup>. These individuals lived in the mid-sixth millennium BC and have significantly more hunter-gatherer-related ancestry than other Balkan Neolithic populations (Fig. 1b, d, Extended Data Figs 1–3, Supplementary Tables 2–4); a model of 82% (confidence interval: 77–86%) northwestern-Anatolian-Neolithic-related, 15% (confidence interval: 12–17%) WHG-related and 4% (confidence interval: 0–9%) EHG-related ancestry fits the data. This hunter-gatherer-related ancestry, with an approximately 4:1 WHG:EHG ratio, plausibly represents a contribution from local Balkan hunter-gatherers genetically similar to those of the Iron Gates region. Late Mesolithic hunter-gatherers in the Balkans were probably





**Figure 3 | Structure and change in northwestern-Anatolian-Neolithic-related populations.** **a**, Populations modelled as a mixture of northwestern Anatolia Neolithic, WHG and EHG. Dashed lines show temporal relationships between populations from the same geographic region. Percentages indicate proportion of WHG + EHG ancestry. Standard errors range from 0.7 to 6.0% (Supplementary Table 3.2.2). **b**, Z scores for the difference between hunter-gatherer-related ancestry on the autosomes and that on the X chromosome, when populations are modelled as a mixture of northwestern Anatolia Neolithic and WHG ( $n = 126$  individuals, group sizes in parentheses). Positive values indicate

more hunter-gatherer-related ancestry on the autosomes, and thus male-biased hunter-gatherer ancestry. ‘Combined’ populations merge all individuals from different times from a geographic area. **c**, Hunter-gatherer-related ancestry proportions on the autosomes, X chromosome (ChrX), mitochondrial DNA (MT; mitochondrial haplogroup U), and the Y chromosome (ChrY; Y chromosome haplogroups I2, R1 and C1). Points show qpAdm (autosomes and X chromosome) or maximum likelihood (mitochondrial DNA and Y chromosome) estimates and bars show approximate 95% confidence intervals ( $n = 109$  individuals, group sizes in parentheses).

concentrated along the coast and major rivers such as the Danube<sup>41</sup>, which directly connects the Iron Gates with Malak Preslavets. Early farmer groups with the highest percentages of hunter-gatherer-related ancestry may, therefore, have been those that lived close to the highest densities of hunter-gatherers.

In the Balkans, Copper Age populations (labelled ‘Balkans\_Chalcolithic’) contain significantly more hunter-gatherer-related ancestry than Neolithic populations as shown, for example, by the statistic  $D(\text{Mbuti}, \text{WHG}, \text{Balkans\_Neolithic}, \text{Balkans\_Chalcolithic})$ ;  $Z = 4.3$  (Supplementary Table 2). This is roughly contemporary with the ‘resurgence’ of hunter-gatherer ancestry previously reported in central Europe and Iberia<sup>7,10,42</sup> and is consistent with changes in funeral rites, specifically the reappearance at around 4500 BC of the Mesolithic tradition of extended supine burial in contrast to the Early Neolithic tradition of flexed burial<sup>43</sup>. Four individuals associated with the Copper Age Trypillian population have approximately 80% northwestern-Anatolian-Neolithic-related ancestry (Supplementary Table 3), confirming that the ancestry of the first farmers of present-day Ukraine was largely derived from the same source as the farmers of Anatolia and western Europe. The roughly 20% of their ancestry from hunter-gatherers is intermediate between WHG and EHG, consistent with it deriving from the Neolithic hunter-gatherers of the region.

We also report genetic data associated with the Late Neolithic Globular Amphora complex. Individuals from two Globular Amphora sites in Poland and Ukraine form a tight cluster, showing high similarity over a large distance (Fig. 1b, d). Both groups of Globular Amphora complex samples had more hunter-gatherer-related ancestry than did Middle Neolithic groups from central Europe<sup>7</sup>; we estimate that the Globular Amphora individuals harboured 25% (confidence interval: 22–27%) WHG-related ancestry, similar to the level seen in Chalcolithic Iberian individuals (Supplementary Table 3). In east-central Europe, the

Globular Amphora complex preceded the Corded Ware complex that marks the appearance of steppe-related ancestry<sup>7,15</sup>, and in southeastern Europe, the Globular Amphora complex bordered populations with steppe-influenced material cultures for hundreds of years<sup>44</sup>. Despite this, the Globular Amphora complex individuals in our study show no evidence of steppe-related ancestry, supporting the hypothesis that this material cultural frontier was also a barrier to gene flow.

About 75% of the ancestry of individuals associated with the Corded Ware complex—and, in central Europe, about 50% of the ancestry of people associated with succeeding archaeological cultures such as the Bell Beaker complex<sup>7,15</sup>—can be traced to a population that probably moved from the Pontic–Caspian steppe in the third millennium BC and that had ancestry similar to that of individuals linked with the Yamnaya complex. However, in two directly dated individuals from southeastern Europe, we find far-earlier evidence of steppe-related ancestry (Fig. 1b, d). One (ANI163) from the Varna I cemetery was dated to 4711–4550 BC and another (I2181) from nearby Smyadovo was dated to 4550–4450 BC. These findings push back the first evidence for steppe-related ancestry this far west in Europe by almost 2,000 years, but it must have been sporadic because other Copper Age (approximately 5000–4000 BC) individuals from the Balkans have no evidence for such ancestry. Bronze Age (approximately 3400–1100 BC) individuals do have steppe-related ancestry: we estimate that they have about 30% (confidence interval: 26–35%), with the highest proportions in the four latest Balkan Bronze Age individuals in our data (later than roughly 1700 BC) and the least in earlier Bronze Age individuals (3400–2500 BC; Fig. 1d).

## Two sources of Asian ancestry in Neolithic Europe

An important question about the initial spread of farming into Europe is whether the first farmers that brought agriculture to northern and

southern Europe were derived from a single source population or instead derived from multiple sources. We confirm that Mediterranean populations, represented in our study by individuals associated with the Epicald Early Neolithic from Iberia<sup>7</sup>, are closely related to Danubian populations represented by the Linearbandkeramik complex from central Europe<sup>7,45</sup>, and that both Mediterranean and Danubian populations are closely related to the Balkan Neolithic population. These three populations form a clade with the northwestern-Anatolian Neolithic individuals as an outgroup, consistent with a single migration into the Balkan Peninsula that then split into two (Supplementary Note 3).

By contrast, data from five southern Greek Neolithic individuals (labelled ‘Peloponnese\_Neolithic’)—three (plus one that has previously been published<sup>26</sup>) from Diros Cave and one from Franchthi Cave—are not consistent with descent from the same source population as other European farmers. *D* statistics (Supplementary Table 2) show that these ‘Peloponnese Neolithic’ individuals, dated to around 4000 BC, are shifted away from WHG, and towards CHG, relative to northwestern-Anatolian Neolithic and Balkan Neolithic individuals. We detect the same pattern in a single Neolithic individual from Krepst in present-day Bulgaria (I0679\_d, 5718–5626 BC). An even more marked shift towards CHG has previously been observed in individuals associated with the Bronze Age Minoan and Mycenaean cultures<sup>26</sup>, suggesting gene flow into the region from populations with CHG-rich ancestry throughout the Neolithic, Chalcolithic and Bronze Age. Possible sources are from people related to the Neolithic population of the central Anatolian site of Tepecik Çiftlik<sup>21</sup>, or the Aegean site of Kumtepe<sup>11</sup>, who are also shifted towards CHG relative to northwestern-Anatolian Neolithic samples, as are later Copper and Bronze Age Anatolians<sup>10,26</sup>.

### Sex bias in Neolithic hunter–gatherer admixture

We provide evidence for sex-biased admixture between hunter-gatherers and farmers in Europe, showing that the Middle Neolithic resurgence of hunter-gatherer-related ancestry<sup>7,42</sup> in central Europe and Iberia was driven more by males than by females (Fig. 3b, c, Extended Data Fig. 5, Supplementary Table 5). To document this, we used qpAdm to compute ancestry proportions on the autosomes and the X chromosome; because males always inherit a maternal X chromosome, differences in these proportions imply sex-biased mixture. There is no evidence of sex bias in the Balkan Neolithic ( $Z = 0.27$ ; a positive  $Z$  score implies bias towards male hunter-gatherer ancestry) or in the Linearbandkeramik and Iberian Early Neolithic ( $Z = -0.22$  and  $1.09$ ). In the Copper Age there is clear bias towards male hunter-gatherer ancestry that is weak in the Balkans ( $Z = 1.66$ ), but stronger in Iberia ( $Z = 3.08$ ) and central Europe ( $Z = 2.74$ ). Consistent with this, hunter-gatherer mitochondrial haplogroups (haplogroup U)<sup>46</sup> are rare and within the intervals of genome-wide ancestry proportions, but hunter-gatherer-associated Y chromosomes<sup>17</sup> are more common (Fig. 3c): seven out of nine male individuals in the Iberian Neolithic and Copper Age and nine out of ten male individuals in Middle–Late Neolithic central Europe (‘Central\_MN’ and ‘Globular\_Amphora’) carried haplogroups I2, R1 or C1.

### No steppe migration to Anatolia via southeast Europe

One version of the steppe hypothesis of Indo-European language origins suggests that Proto-Indo-European languages developed north of the Black and Caspian seas, and that the earliest-known diverging branch, the Anatolian branch, was spread into Asia Minor by the movements of steppe peoples through the Balkan Peninsula during the Copper Age at around 4000 BC<sup>47</sup>. If this were correct, then one way to detect evidence of the spread of Indo-European languages would be the appearance of large amounts of steppe-related ancestry first in the Balkan Peninsula, and later in Anatolia. However, our data provide no evidence for this scenario. Although we find sporadic steppe-related ancestry in Balkan Copper and Bronze Age individuals, this ancestry is rare until the late Bronze Age. Furthermore, although Bronze Age Anatolian individuals have CHG-related ancestry<sup>26</sup>, they

do not have the EHG-related ancestry characteristic of all steppe populations sampled to date<sup>19</sup> or the WHG-related ancestry that is ubiquitous in Neolithic southeastern Europe (Extended Data Figs 2, 3, Supplementary Table 2). We caution, however, that at present we only have data from a small number of Bronze Age Anatolian individuals, none of whom are associated with known Indo-European-speaking populations. An alternative hypothesis is that the homeland of Proto-Indo-European languages was in the Caucasus or in Iran. In this scenario, westward population movement contributed to the dispersal of Anatolian languages, and northward movement and mixture with EHG was responsible for the formation of a ‘Late Proto-Indo-European’-speaking population associated with the Yamnaya complex<sup>13</sup>. Although this scenario gains plausibility from our results, it remains possible that Indo-European languages were spread through southeastern Europe into Anatolia without large-scale population movement or admixture.

### Discussion

Our study shows that southeastern Europe served as a genetic contact zone between east and west over thousands of years. Before the arrival of farming, the region saw interaction between diverged groups of hunter-gatherers, and this interaction continued after farming arrived. Although this study has clarified the genomic history of the region from the Mesolithic to the Bronze Age, the processes that connected these populations to ones living today remain largely unknown. An important priority for future research should be to sample populations from the Bronze Age, Iron Age, Roman and Medieval periods and to compare them to present-day populations to understand how these transitions occurred.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 6 May 2017; accepted 16 January 2018.**

**Published online 21 February 2018.**

1. Tringham, R. E. in *The Transition to Agriculture in Prehistoric Europe* (ed. Price, D.) 19–56 (Cambridge Univ. Press, 2000).
2. Bellwood, P. *First Farmers: The Origins of Agricultural Societies* 2nd edn (Wiley–Blackwell, 2004).
3. Golitko, M. in *Ancient Europe, 8000 B.C. to A.D. 1000: An Encyclopedia of the Barbarian World* (eds Bogucki, P. & Crabtree, P. J.) 259–266 (Charles Scribners & Sons, 2003).
4. VanderLinden, M. in *Investigating Archaeological Cultures: Material Culture, Variability, and Transmission* (eds Roberts, B. W. & VanderLinden, M.) 289–319 (Springer, 2012).
5. Bramanti, B. et al. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326**, 137–140 (2009).
6. Skoglund, P. et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**, 466–469 (2012).
7. Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
8. Cassidy, L. M. et al. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc. Natl Acad. Sci. USA* **113**, 368–373 (2016).
9. Hofmanová, Z. et al. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl Acad. Sci. USA* **113**, 6886–6891 (2016).
10. Mathieson, I. et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
11. Omrak, A. et al. Genomic evidence establishes Anatolia as the source of the European Neolithic gene pool. *Curr. Biol.* **26**, 270–275 (2016).
12. Müller, J., Rassmann, K. & Videiko, M. *Trypillia Mega-Sites and European Prehistory: 4100–3400 BCE* (Routledge, 2016).
13. Anthony, D. W. *The Horse, the Wheel and Language* (Princeton Univ. Press, 2007).
14. Fu, Q. et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
15. Allentoft, M. E. et al. Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
16. Fu, Q. et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
17. Fu, Q. et al. The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).
18. Gallego Llorente, M. et al. Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science* **350**, 820–822 (2015).

19. Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015).
20. Keller, A. *et al.* New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **3**, 698 (2012).
21. Kiling, G. M. *et al.* The demographic development of the first farmers in Anatolia. *Curr. Biol.* **26**, 2659–2666 (2016).
22. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
23. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
24. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–228 (2014).
25. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).
26. Lazaridis, I. *et al.* Genetic origins of the Minoans and Mycenaeans. *Nature* **548**, 214–218 (2017).
27. Lipson, M. *et al.* Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature* **551**, 368–372 (2017).
28. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
29. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
30. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
31. Gronenborn, D. & Dolukhanov, P. In *The Oxford Handbook of Neolithic Europe* (eds Fowler, C., Harding, J. & Hofmann, D.) 195–214 (Oxford Univ. Press, 2015).
32. Telegin, D. Y. Neolithic cultures of Ukraine and adjacent areas and their chronology. *J. World Prehist.* **1**, 307–331 (1987).
33. Telegin, D. Y. & Potekhina, I. D. *Neolithic Cemeteries and Populations in the Dnieper Basin* (British Archaeological Reports, 1987).
34. Jones, E. R. *et al.* The Neolithic transition in the Baltic was not driven by admixture with early European farmers. *Curr. Biol.* **27**, 576–582 (2017).
35. Mitnik, A. *et al.* The genetic prehistory of the Baltic Sea region. *Nat. Commun.* **9**, 443 (2018).
36. Saag, L. *et al.* Extensive farming in Estonia started through a sex-biased migration from the Steppe. *Curr. Biol.* **27**, 2185–2193.e6 (2017).
37. Maier, A. *The Central European Magdalenian: Regional Diversity and Internal Variability* (Springer, 2015).
38. Boric, D. & Price, T. D. Strontium isotopes document greater human mobility at the start of the Balkan Neolithic. *Proc. Natl Acad. Sci. USA* **110**, 3298–3303 (2013).
39. Krauß, R., Marinova, E., De Brue, H. & Weninger, B. The rapid spread of early farming from the Aegean into the Balkans via the Sub-Mediterranean–Aegean vegetation Zone. *Quat. Int.* <https://doi.org/10.1016/j.quaint.2017.01.019> (2017).
40. Bacvarov, K. In *Moments in Time: Papers Presented to Pál Raczky on his 60th Birthday* (eds Anders, A. & Kulcsár, G.) 29–34 (L'Harmattan, 2013).
41. Gurova, M. & Bonsall, C. 'Pre-Neolithic' in Southeast Europe: a Bulgarian perspective. *Documenta Praehistorica* **41**, 95–109 (2014).
42. Brandt, G. *et al.* Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science* **342**, 257–261 (2013).
43. Boric, D. In *The Oxford Handbook of Neolithic Europe* (eds Fowler, C., Harding, J. & Hofmann, D.) 927–957 (Oxford Univ. Press, 2015).
44. Szmtyt, M. In *Transition to the Bronze Age (Archaeolingua 30)* (eds Heyd, V., Kulcsár, G. & Szeverényi, V.) 93–111 (Archaeolingua, 2013).
45. Olalde, I. *et al.* A common genetic origin for early farmers from Mediterranean Cardial and central European LBK cultures. *Mol. Biol. Evol.* **32**, 3132–3142 (2015).
46. Posth, C. *et al.* Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe. *Curr. Biol.* **26**, 827–833 (2016).
47. Anthony, D. W. & Ringe, D. The Indo-European homeland from linguistic and archaeological perspectives. *Annu. Rev. Linguist.* **1**, 199–219 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank D. Anthony, I. Lazaridis and M. Lipson for comments on the manuscript, B. Llamas, A. Cooper and A. Furtwängler for contributions to laboratory work, R. Evershed for contributing <sup>14</sup>C dates and F. Novotny for assistance with samples. Support for this project was provided by the Human Frontier Science Program fellowship LT001095/2014-L to I.M., by DFG grant AL 287 / 14-1 to K.W.A.; by Irish Research Council grant GOIPG/2013/36 to D.F.; by the NSF Archaeometry program BCS-1460369 to D.J.K. for AMS <sup>14</sup>C work; by MEN-UEFISCDI grant, Partnerships in Priority Areas Program – PN II (PN-II-PT-PCCA-2013-4-2302) to C.L.; by Croatian Science Foundation grant IP-2016-06-1450 to M.N. and I.J.; by European Research Council grant ERC CoG 724703 and Deutsche Forschungsgemeinschaft DFG FOR2237 to K.H.; by ERC starting grant ADNABIOARC (263441) to R.P.; and by US National Science Foundation HOMOINID grant BCS-1032255, US National Institutes of Health grant GM100233, the Howard Hughes Medical Institute and an Allen Discovery Center grant from the Paul Allen Foundation to D.R.

**Author Contributions** S.A.-R., A.S.-N., S.Vai., S.A., K.W.A., R.A., D.A., A.A., N.A., K.B., M.B.G., H.B., M.B., A.Bo., Y.B., A.Bu., J.B., S.C., N.J.C., R.C., M.C., C.C., D.G.D., N.E., M.Fr., B.Gal., G.G., B.Ge., T.Ha., V.H., K.H., T.Hi., S.I., I.J., I.Ka., D.Ko., A.K., D.La., M.La., C.L., M.Le., K.L., D.L.V., D.Lo., I.L., M.Ma., F.M., K.M., H.M., M.Me., P.M., V.M., V.P., T.D.P., A.Si., L.S., M.S., V.S., P.S., A.St., T.S., M.T.-N., C.T., I.V., F.Va., S.Vas., F.Ve., S.Ve., E.V., B.V., C.V., J.Z., S.Z., P.W.S., G.C., R.K., D.C., G.Z., B.Gay., M.Li., A.G.N., I.P., A.P., D.B., C.B., J.K., R.P. and D.R. assembled and interpreted archaeological material. C.P., A.S.-N., N.R., N.B., F.C., O.C., D.F., M.Fe., B.Gam., G.G.F., W.H., E.H., E.J., D.Ke., B.K.-K., I.Ku., M.Mi., A.M., K.N., M.N., J.O., S.P., K.Si., K.St. and S.Vai. performed laboratory work. I.M., C.P., A.S.-N., S.M., I.O., N.P. and D.R. analysed data. D.J.K., S.T., D.B. and C.B. interpreted <sup>14</sup>C dates. J.K., R.P. and D.R. supervised analysis or laboratory work. I.M. and D.R. wrote the paper with input from all co-authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.R. ([reich@genetics.med.harvard.edu](mailto:reich@genetics.med.harvard.edu)), R.P. ([ron.pinhasi@univie.ac.at](mailto:ron.pinhasi@univie.ac.at)) or I.M. ([mathi@pennmedicine.upenn.edu](mailto:mathi@pennmedicine.upenn.edu)).

**Reviewer Information** Nature thanks C. Renfrew, A. Scally and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Iain Mathieson<sup>1†</sup>, Songül Alpaslan-Roodenberg<sup>1</sup>, Cosimo Posth<sup>2,3</sup>, Anna Szécsényi-Nagy<sup>4</sup>, Nadin Rohland<sup>1</sup>, Swapan Mallick<sup>1,5</sup>, Iñigo Olalde<sup>1</sup>, Nasreen Broomandkhoshbacht<sup>1,5</sup>, Francesca Candilio<sup>6</sup>, Olivia Cheronet<sup>6,7</sup>, Daniel Fernandes<sup>6,8</sup>, Daniel Ferry<sup>1,5</sup>, Beatriz Gamarra<sup>6</sup>, Gloria González Fortes<sup>9</sup>, Wolfgang Haak<sup>2,10</sup>, Eadaoin Harney<sup>1,5</sup>, Eppie Jones<sup>11,12</sup>, Denise Keating<sup>6</sup>, Ben Krause-Kyora<sup>2</sup>, Isil Kucukkalipci<sup>3</sup>, Megan Michel<sup>1,5</sup>, Alissa Mitnik<sup>2,3</sup>, Kathrin Nägele<sup>2</sup>, Mario Novak<sup>6,13</sup>, Jonas Oppenheimer<sup>1,5</sup>, Nick Patterson<sup>14</sup>, Saskia Pfrengle<sup>3</sup>, Kendra Sirak<sup>6,15</sup>, Kristin Stewardson<sup>1,5</sup>, Stefania Vai<sup>16</sup>, Stefan Alexandrov<sup>17</sup>, Kurt W. Alt<sup>18,19,20</sup>, Radian Andreescu<sup>21</sup>, Dragana Antonović<sup>22</sup>, Abigail Ash<sup>6</sup>, Nadezhda Atanassova<sup>23</sup>, Krum Bacvarov<sup>17</sup>, Mende Balázs Gusztáv<sup>4</sup>, Hervé Bocherens<sup>24,25</sup>, Michael Bolus<sup>26</sup>, Adina Boronean<sup>27</sup>, Yavor Boyadzhiev<sup>17</sup>, Alicia Budnik<sup>28</sup>, Josip Burmaz<sup>29</sup>, Stefan Chohadzhiev<sup>30</sup>, Nicholas J. Conard<sup>25,31</sup>, Richard Cottiaux<sup>32</sup>, Maja Čuka<sup>33</sup>, Christophe Cupillard<sup>34,35</sup>, Dorothee G. Drucker<sup>25</sup>, Nedko Elenski<sup>36</sup>, Michael Francken<sup>37</sup>, Borislava Galabova<sup>38</sup>, Georgi Ganetsovski<sup>39</sup>, Bernard Gély<sup>40</sup>, Tamás Hajdu<sup>41</sup>, Veneta Handzhyska<sup>42</sup>, Katerina Harvati<sup>25,37</sup>, Thomas Higham<sup>43</sup>, Stanislav Iliev<sup>44</sup>, Ivor Janković<sup>3,45</sup>, Ivor Karavanić<sup>45,46</sup>, Douglas J. Kennet<sup>47</sup>, Darko Komšo<sup>33</sup>, Alexandra Kozak<sup>48</sup>, Damian Labuda<sup>49</sup>, Martina Lari<sup>16</sup>, Catalin Lazar<sup>21,50</sup>, Maleen Leppek<sup>51</sup>, Krassimir Leshtakov<sup>42</sup>, Domenico Lo Vetro<sup>52,53</sup>, Dženi Los<sup>29</sup>, Ivalyo Lozanov<sup>42</sup>, Maria Malina<sup>26</sup>, Fabio Martini<sup>52,53</sup>, Kath McSweeney<sup>54</sup>, Harald Meller<sup>20</sup>, Marko Mendišić<sup>55</sup>, Pavel Mirea<sup>56</sup>, Vyacheslav Moiseyev<sup>57</sup>, Vanya Petrova<sup>42</sup>, T. Douglas Price<sup>58</sup>, Angela Simalcsik<sup>59</sup>, Luca Sineo<sup>60</sup>, Mario Šlaus<sup>61</sup>, Vladimir Slavchev<sup>62</sup>, Petar Stanev<sup>63</sup>, Andrej Starović<sup>63</sup>, Tamás Szeniczey<sup>41</sup>, Sahra Talamo<sup>64</sup>, Maria Teschler-Nicola<sup>7,65</sup>, Corinne Thevenet<sup>66</sup>, Ivan Valchev<sup>42</sup>, Frédéricque Valentin<sup>67</sup>, Sergey Vasilyev<sup>68</sup>, Fanica Veljanovska<sup>69</sup>, Svetlana Venelinova<sup>70</sup>, Elizaveta Veselovskaya<sup>68</sup>, Bence Viola<sup>71,72</sup>, Cristian Virag<sup>73</sup>, Joško Zaninović<sup>74</sup>, Steve Zäuner<sup>75</sup>, Philipp W. Stockhammer<sup>2,51</sup>, Giulio Catalano<sup>60</sup>, Raiko Krauß<sup>76</sup>, David Caramelli<sup>16</sup>, Gunita Zariņa<sup>77</sup>, Bissiera Gaydarska<sup>78</sup>, Malcolm Lillie<sup>79</sup>, Alexey G. Nikitin<sup>80</sup>, Inna Potekhina<sup>48</sup>, Anastasia Papathanasiou<sup>81</sup>, Dušan Boric<sup>82</sup>, Clive Bonsall<sup>54</sup>, Johannes Krause<sup>2,3</sup>, Ron Pinhasi<sup>6,7,\*</sup> & David Reich<sup>1,5,14,\*</sup>

<sup>1</sup>Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

<sup>2</sup>Department of Archaeogenetics, Max Planck Institute for the Science of Human History, 07745 Jena, Germany. <sup>3</sup>Institute for Archaeological Sciences, University of Tübingen, Tübingen, Germany. <sup>4</sup>Laboratory of Archaeogenetics, Institute of Archaeology, Research Centre for the Humanities, Hungarian Academy of Sciences, H-1097 Budapest, Hungary. <sup>5</sup>Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>6</sup>Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland. <sup>7</sup>Department of Anthropology, University of Vienna, 1090 Vienna, Austria. <sup>8</sup>CIA, Department of Life Sciences, University of Coimbra, 3000-456 Coimbra, Portugal. <sup>9</sup>Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara 44100, Italy. <sup>10</sup>Australian Centre for Ancient DNA, School of Biological Sciences, The University of Adelaide, SA-5005 Adelaide, South Australia, Australia. <sup>11</sup>Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland. <sup>12</sup>Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK. <sup>13</sup>Institute for Anthropological Research, 10000 Zagreb, Croatia. <sup>14</sup>Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.

<sup>15</sup>Department of Anthropology, Emory University, Atlanta, Georgia 30322, USA. <sup>16</sup>Dipartimento di Biologia, Università di Firenze, 50122 Florence, Italy. <sup>17</sup>National Institute of Archaeology and Museum, Bulgarian Academy of Sciences, BG-1000 Sofia, Bulgaria. <sup>18</sup>Danube Private University, A-3500 Krems, Austria. <sup>19</sup>Department of Biomedical Engineering and Integrative Prehistory and Archaeological Science, CH-4123 Basel-Allschwil, Switzerland. <sup>20</sup>State Office for Heritage Management and Archaeology Saxony-Anhalt and State Museum of Prehistory, 06114 Halle, Germany. <sup>21</sup>National History Museum of Romania, 030026, Bucharest, Romania. <sup>22</sup>Institute of Archaeology, Belgrade, Serbia. <sup>23</sup>Institute of Experimental Morphology, Pathology and Anthropology with Museum, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria.



- <sup>24</sup>Department of Geosciences, Biogeology, Universität Tübingen, 72074 Tübingen, Germany.
- <sup>25</sup>Senckenberg Centre for Human Evolution and Palaeoenvironment at the University of Tübingen, 72076 Tübingen, Germany.
- <sup>26</sup>ROCEEH Research Center, Heidelberg Academy of Sciences and Humanities, University of Tübingen, 72070 Tübingen, Germany.
- <sup>27</sup>Vasile Pârvan Institute of Archaeology, Romanian Academy, 010667 Bucharest, Romania.
- <sup>28</sup>Human Biology Department, Cardinal Stefan Wyszyński University, 01-938 Warsaw, Poland.
- <sup>29</sup>KADUCEJ d.o.o., 21000 Split, Croatia.
- <sup>30</sup>St. Cyril and Methodius University, 5000 Veliko Turnovo, Bulgaria.
- <sup>31</sup>Department of Early Prehistory and Quaternary Ecology, University of Tübingen, 72070 Tübingen, Germany.
- <sup>32</sup>INRAP/UMR 8215 Trajectoires, 92023 Nanterre, France.
- <sup>33</sup>Archaeological Museum of Istria, 52100 Pula, Croatia.
- <sup>34</sup>Service Régional de l'Archéologie de Bourgogne-Franche-Comté, 25043 Besançon Cedex, France.
- <sup>35</sup>Laboratoire Chronoenvironnement, UMR 6249 du CNRS, UFR des Sciences et Techniques, 25030 Besançon Cedex, France.
- <sup>36</sup>Regional Museum of History Veliko Tarnovo, 5000 Veliko Tarnovo, Bulgaria.
- <sup>37</sup>Institute for Archaeological Sciences, Paleoanthropology, University of Tübingen, 72070 Tübingen, Germany.
- <sup>38</sup>Laboratory for Human Bioarchaeology, 1202 Sofia, Bulgaria.
- <sup>39</sup>Regional Museum of History, 3000 Vratsa, Bulgaria.
- <sup>40</sup>DRAC Auvergne - Rhône Alpes, Ministère de la Culture, Lyon Cedex 01, France.
- <sup>41</sup>Eötvös Loránd University, Faculty of Science, Institute of Biology, Department of Biological Anthropology, H-1117 Budapest, Hungary.
- <sup>42</sup>Department of Archaeology, Sofia University St. Kliment Ohridski, 1504 Sofia, Bulgaria.
- <sup>43</sup>Oxford Radiocarbon Accelerator Unit, Research Laboratory for Archaeology and the History of Art, University of Oxford, Dyson Perrins Building, Oxford OX1 3QY, UK.
- <sup>44</sup>Regional Museum of History, 6300 Haskovo, Bulgaria.
- <sup>45</sup>Department of Anthropology, University of Wyoming, Laramie, Wyoming 82071, USA.
- <sup>46</sup>Department of Archaeology, Faculty of Humanities and Social Sciences, University of Zagreb, 10000 Zagreb, Croatia.
- <sup>47</sup>Department of Anthropology and Institutes for Energy and the Environment, Pennsylvania State University, University Park, Pennsylvania 16802, USA.
- <sup>48</sup>Department of Bioarchaeology, Institute of Archaeology, National Academy of Sciences of Ukraine, 04210 Kiev, Ukraine.
- <sup>49</sup>CHU Sainte-Justine Research Center, Pediatric Department, Université de Montréal, Montreal, Québec H3T 1C5, Canada.
- <sup>50</sup>Department of Ancient History, Archaeology and History of Art, Faculty of History, University of Bucharest, 50107 Bucharest, Romania.
- <sup>51</sup>Institute for Pre- and Protohistoric Archaeology and the Archaeology of the Roman Provinces, Ludwig-Maximilians-University, 80799 Munich, Germany.
- <sup>52</sup>Dipartimento SAGAS - Sezione di Archeologia e Antico Oriente, Università degli Studi di Firenze, 50122 Florence, Italy.
- <sup>53</sup>Museo e Istituto fiorentino di Preistoria, 50122 Florence, Italy.
- <sup>54</sup>School of History, Classics and Archaeology, University of Edinburgh, Edinburgh EH8 9AG, UK.
- <sup>55</sup>Conservation Department in Šibenik, Ministry of Culture of the Republic of Croatia, 22000 Šibenik, Croatia.
- <sup>56</sup>Teleorman County Museum, 140033 Alexandria, Romania.
- <sup>57</sup>Peter the Great Museum of Anthropology and Ethnography (Kunstkamera) RAS, 199034 St. Petersburg, Russia.
- <sup>58</sup>Department of Anthropology, University of Wisconsin, Madison, Wisconsin 53706, USA.
- <sup>59</sup>Olga Necrasov Centre for Anthropological Research, Romanian Academy – Iași Branch, 700481 Iași, Romania.
- <sup>60</sup>Dipartimento di Scienze e tecnologie biologiche, chimiche e farmaceutiche, Lab. of Anthropology, Università degli studi di Palermo, 90133 Palermo, Italy.
- <sup>61</sup>Anthropological Center, Croatian Academy of Sciences and Arts, 10000 Zagreb, Croatia.
- <sup>62</sup>Regional Historical Museum Varna, BG-9000 Varna, Bulgaria.
- <sup>63</sup>National Museum in Belgrade, Belgrade, Serbia.
- <sup>64</sup>Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany.
- <sup>65</sup>Department of Anthropology, Natural History Museum Vienna, 1010 Vienna, Austria.
- <sup>66</sup>INRAP/UMR 8215 Trajectoires, 92023 Nanterre, France.
- <sup>67</sup>CNRS/UMR 7041 ArScAn MAE, 92023 Nanterre, France.
- <sup>68</sup>Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow, 119991, Russia.
- <sup>69</sup>Archaeological Museum of Macedonia, 1000 Skopje, the former Yugoslav Republic of Macedonia.
- <sup>70</sup>Regional Museum of History, 9700 Shumen, Bulgaria.
- <sup>71</sup>Department of Anthropology, University of Toronto, Toronto, Ontario, M5S 2S2, Canada.
- <sup>72</sup>Institute of Archaeology & Ethnography, Siberian Branch, Russian Academy of Sciences, Novosibirsk 630090, Russia.
- <sup>73</sup>Satu Mare County Museum Archaeology Department, 440026 Satu Mare, Romania.
- <sup>74</sup>Municipal Museum Drniš, 22320 Drniš, Croatia.
- <sup>75</sup>anthropol - Anthropologieservice, 72379 Hechingen, Germany.
- <sup>76</sup>Institute for Prehistory, Early History and Medieval Archaeology, University of Tübingen, 72070 Tübingen, Germany.
- <sup>77</sup>Institute of Latvian History, University of Latvia, Rīga 1050, Latvia.
- <sup>78</sup>Department of Archaeology, Durham University, Durham DH1 3LE, UK.
- <sup>79</sup>School of Environmental Sciences, Geography, University of Hull, Hull HU6 7RX, UK.
- <sup>80</sup>Department of Biology, Grand Valley State University, Allendale, Michigan 49401, USA.
- <sup>81</sup>Ephorate of Paleoanthropology and Speleology, 11636 Athens, Greece.
- <sup>82</sup>The Italian Academy for Advanced Studies in America, Columbia University, New York, New York 10027, USA.
- †Present address: Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

\*These authors contributed equally to this work.

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Ancient DNA analysis.** We extracted DNA and prepared next-generation sequencing libraries in four different dedicated ancient DNA laboratories (Adelaide, Boston, Budapest and Tübingen). We also prepared powder in a fifth laboratory (Dublin), which was sent to Boston for DNA extraction and library preparation (Supplementary Table 1).

Two samples were processed at the Australian Centre for Ancient DNA, Adelaide, according to previously published methods<sup>7</sup> and sent to Boston for sub-sequencing, 1240k capture and sequencing.

Seven samples were processed<sup>27</sup> at the Institute of Archaeology RCH HAS, Budapest, and amplified libraries were sent to Boston for screening, 1240k capture and sequencing.

Seventeen samples were processed at the Institute for Archaeological Sciences of the University of Tübingen and at the Max Planck Institute for the Science of Human History in Jena. Extraction<sup>48</sup> and library preparation<sup>49,50</sup> followed established protocols. We performed in-solution enrichment for sequences overlapping about 1.24 million SNPs ('1240k capture'), and sequenced on an Illumina HiSeq 4000 or NextSeq 500 for 76 bp using either single- or paired-end sequencing.

The remaining 199 samples were processed at Harvard Medical School, Boston. Using about 75 mg of sample powder from each sample (prepared from skeletal samples either in Boston or University College Dublin, Dublin), we extracted DNA following established methods<sup>48</sup> replacing the column assembly with the column extenders from a Roche kit<sup>51</sup>. We prepared double barcoded libraries with truncated adapters from between one ninth and one third of the DNA extract. Most libraries included in the nuclear genome analysis (90%) were subjected to partial ('half') uracil-DNA-glycosylase (UDG) treatment before blunt-end repair. This treatment reduces by an order of magnitude the characteristic cytosine-to-thymine errors of ancient DNA data<sup>52</sup> but works inefficiently at the 5' ends<sup>50</sup>, thereby leaving a signal of characteristic damage at the terminal ends of ancient sequences. Some libraries were not UDG-treated ('minus'). For a subset of samples, we increased coverage by preparing additional libraries from the existing DNA extract using partial UDG library preparation, but replacing the MinElute column cleanups in between enzymatic reactions with magnetic bead cleanups, and the final PCR cleanup with SPRI bead cleanup<sup>53,54</sup>. We screened all libraries from Adelaide, Boston and Budapest by enriching for the mitochondrial genome, plus about 3,000 (50 in an earlier and unpublished version) nuclear SNPs using a bead-based capture<sup>55</sup> but with the probes replaced by amplified oligonucleotides synthesized by CustomArray. After the capture, we completed the adapter sites using PCR, attaching dual index combinations<sup>56</sup> to each enriched library. We sequenced the products of between 100 and 200 libraries together with the non-enriched libraries (shotgun) on an Illumina NextSeq500 using v.2 150 cycle kits for 2 × 76 cycles and 2 × 7 cycles.

In Boston, we performed two rounds of in-solution enrichment ('1240k capture') for a targeted set of 1,237,207 SNPs using previously reported protocols<sup>7,14,23</sup>. For a total of 41 individuals, we increased coverage by building one to nine additional libraries for the same sample. When we built multiple libraries from the same extract, we often pooled them in equimolar ratios before the capture. We performed all sequencing on an Illumina NextSeq500 using v.2 150 cycle kits for 2 × 76 cycles and 2 × 7 cycles. We attempted to sequence each enriched library up to the point at which we estimated that it was economically inefficient to sequence further. Specifically, we iteratively sequenced more from each individual and only stopped when we estimated that the expected increase in the number of targeted SNPs hit at least once would be less than about one for every 100 new read pairs generated. After sequencing and merging the paired reads into a single sequence, we trimmed two bases from the end of each sequence and aligned to the human genome (b37/hg19) using bwa<sup>57</sup>. We then removed individuals with evidence of contamination based on mitochondrial DNA polymorphism<sup>58</sup> or difference in principal component analysis space between damaged and undamaged reads<sup>59</sup>, a high rate of heterozygosity on the chromosome X despite being male<sup>59,60</sup> or a ratio of X-to-Y chromosome sequences different than would be expected from either an XX or XY karyotype. We also removed individuals that had low coverage (fewer than 15,000 SNPs hit on the autosomes). We report, but do not analyse, data from nine individuals that were first-degree relatives of others in the dataset (determined by comparing rates of allele sharing between pairs of individuals).

After removing a small number of sites that failed to capture, we were left with a total of 1,233,013 sites of which 32,670 were on chromosome X and 49,704 were on chromosome Y, with a median coverage at targeted SNPs on the 215 newly reported individuals of 0.90 (range 0.007–9.2; Supplementary Table 1). We generated 'pseudo-haploid' calls by selecting a single sequence randomly for

each individual at each SNP. Thus, there is only a single allele from each individual at each site, but adjacent alleles might come from either of the two haplotypes of the individual. We merged the newly reported data with previously reported data from 274 other ancient individuals<sup>9–11,15–27</sup>, making pseudo-haploid calls in the same way at the 1240k sites for individuals that were shotgun-sequenced rather than captured.

Using the captured mitochondrial sequence from the screening process, we called mitochondrial haplogroups. Using the SNPs on the Y chromosome, we called Y chromosome haplogroups for males by restricting to sequences with mapping quality  $\geq 30$  and bases with base quality  $\geq 30$ . We determined the most derived mutation for each individual, using the nomenclature of the International Society of Genetic Genealogy (<http://www.isogg.org>) version 11.110 (accessed 21 April 2016).

**Population genetic analysis.** To analyse these ancient individuals in the context of present-day genetic diversity, we merged them with the following two datasets: (1) 300 high-coverage genomes from a diverse worldwide set of 142 populations sequenced as part of the Simons Genome Diversity Project<sup>28</sup> ('SGDP merge') and (2) 777 west Eurasian individuals genotyped on the Human Origins array<sup>23</sup>, with 597,573 sites in the merged dataset ('HO merge').

We computed principal components of the present-day individuals in the HO merge and projected the ancient individuals onto the first two components using the 'lsqproject: YES' option in smartpca (v.15100)<sup>61</sup> (<https://www.hsph.harvard.edu/alkes-price/software/>).

We ran ADMIXTURE (v.1.3.0) in both supervised and unsupervised mode. In supervised mode we used only the ancient individuals, on the full set of SNPs, and with the following population labels fixed: Anatolia\_Neolithic, WHG, EHG, and Yamnaya.

For unsupervised mode we used the HO merge dataset, including 777 present-day individuals. We flagged individuals that were genetic outliers based on principal components analysis and ADMIXTURE relative to other individuals from the same time period and archaeological culture.

We computed *D* statistics using qpDstat (v.710). *D* statistics of the form *D*(*A*,*B*,*X*,*Y*) test the null hypothesis of the unrooted tree topology ((*A*,*B*),(*X*,*Y*)). A positive value indicates that either *A* and *X*, or *B* and *Y*, share more drift than expected under the null hypothesis. We quote *D* statistics as the *Z* score computed using default block jackknife parameters.

We fitted admixture proportions with qpAdm (v.610) using the SGDP merge. Given a set of outgroup ('right') populations, qpAdm models one of a set of source ('left') populations (the 'test' population) as a mixture of the other sources by fitting admixture proportions to match the observed matrix of *f<sub>i</sub>* statistics as closely as possible. We report a *P* value for the null hypothesis that the test population does not have ancestry from another source that is differentially related to the right populations. We computed standard errors for the mixture proportions using a block jackknife. Importantly, qpAdm does not require that the source populations are actually the admixing populations—nor does it require that they are unadmixed themselves. Instead, qpAdm only requires that source populations are a clade with the correct admixing populations, relative to the other sources. Infeasible coefficient estimates (that is, outside [0, 1]) are usually a sign of poor model fit, but in the case where the source with a negative coefficient is itself admixed, could be interpreted as implying that the true source is a population with different admixture proportions. We used the following set of seven populations as outgroups or 'right populations': Mbuti.DG, Ust\_Ishim\_HG\_published.DG, Mota.SG, MA1\_HG.SG, Villabruna, Papuan.DG, Onge.DG and Han.DG.

For some analyses for which we required extra resolution (Supplementary Table 4) we used an extended set of 14 right (outgroup) populations, including additional Upper Palaeolithic European individuals<sup>17</sup>: ElMiron, Mota.SG, Mbuti.DG, Ust\_Ishim\_HG\_published.DG, MA1\_HG.SG, AfontovaGora3, GoyetQ116-1\_published, Villabruna, Kostenki14, Vestonice16, Karitiana.DG, Papuan.DG, Onge.DG and Han.DG.

We also fitted admixture graphs with qpGraph (v.6021)<sup>30</sup> (<https://github.com/DReichLab/AdmixTools>, Supplementary Note 3). As with qpAdm, qpGraph also tries to match a matrix of *f* statistics, but rather than fitting one population as a mixture of other, specified, populations, it fits the relationship between all tested populations simultaneously, potentially incorporating multiple admixture events. However, qpGraph requires the graph relating the populations to be specified in advance. We tested goodness-of-fit by computing the expected *D* statistics under the fitted model, finding the largest *D* statistic outlier between the fitted and observed model, and then computing a *Z* score using a block jackknife.

For 114 individuals with hunter-gatherer-related ancestry, we estimated an effective migration surface using the software EEMS (<https://github.com/dipetkov/eems>)<sup>62</sup>. We computed pairwise differences between individuals using the bed2diffs2 program provided with EEMS. We set the number of demes to 400 and

defined the outer boundary of the region by the polygon (in latitude–longitude coordinates) ((66,60), (60,10), (45,−15), (35,−10), (35,60)). We ran the Markov chain Monte Carlo ten times with different random seeds, each time with one million burn-in and four million regular iterations, thinned to one in ten thousand.

To analyse potential sex bias in admixture, we used qpAdm to estimate admixture proportions on the autosomes (default option) and on the X chromosome (option “chrom: 23”). We computed  $Z$  scores for the difference between the autosomes and the X chromosome as  $Z = \frac{p_A - p_X}{\sqrt{\sigma_A^2 + \sigma_X^2}}$  in which  $p_A$  and  $p_X$  are the

hunter-gatherer admixture proportions on the autosomes and the X chromosome, and  $\sigma_A$  and  $\sigma_X$  are the corresponding jackknife standard deviations. Thus, a positive  $Z$  score means that there is more hunter-gatherer admixture on the autosomes than on the X chromosome, indicating that the hunter-gatherer admixture was male-biased. Because X chromosome standard errors are high and qpAdm results can be sensitive to which population is first in the list of outgroup populations, we checked that the patterns we observed were robust to cyclic permutation of the outgroups. To compare frequencies of hunter-gatherer uniparental markers, we counted the individuals with mitochondrial haplogroup U and Y chromosome haplogroups C1, I2 and R1, which are all common in Mesolithic hunter-gatherers but rare or absent in northwestern-Anatolian Neolithic individuals. The Iron Gates hunter-gatherers also carry H and K1 mitochondrial haplogroups, so the proportion of haplogroup U represents the minimum maternal hunter-gatherer contribution. We computed binomial confidence intervals for the proportion of haplogroups associated with each ancestry type using the Agresti–Coull method<sup>63,64</sup> implemented in the binom package in R.

Given autosomal and X chromosome admixture proportions, we estimated the proportion of male and female hunter-gatherer ancestors by assuming a single-pulse model of admixture. If the proportions of male and female ancestors that are hunter-gatherer-related are given by  $m$  and  $f$ , respectively, then the proportions of hunter-gatherer-related ancestry on the autosomes and the X chromosome are given by  $\frac{m+f}{2}$  and  $\frac{m+2f}{3}$ . We approximated the sampling error in the observed admixture proportions by the estimated jackknife error and computed the likelihood surface for  $(m,f)$  over a grid ranging from (0,0) to (1,1).

**Direct AMS  $^{14}\text{C}$  bone dates.** We report 137 direct AMS  $^{14}\text{C}$  bone dates for 136 individuals from multiple AMS radiocarbon laboratories. In general, bone samples were manually cleaned and demineralized in weak HCl and, in most cases (Radiocarbon laboratory codes: UCIAMS and OxA), soaked in an alkali bath (NaOH) at room temperature to remove contaminating soil humates. Samples were then rinsed to neutrality in Nanopure  $\text{H}_2\text{O}$  and gelatinized in HCl<sup>65</sup>. The resulting gelatin was lyophilized and weighed to determine per cent yield as a measure of collagen preservation (percentage crude gelatin yield). Collagen was then directly AMS  $^{14}\text{C}$ -dated (Radiocarbon laboratory codes: Beta, AA) or further purified using ultrafiltration (Radiocarbon laboratory codes: PSU, UCIAMS, OxA, Poz, and MAMS)<sup>66</sup>. It is standard in some laboratories (Radiocarbon laboratory codes: PSU, UCIAMS and OxA) to use stable carbon and nitrogen isotopes as an additional quality control measure. For these samples, the percentage of C, percentage of N and C:N ratios were evaluated before AMS  $^{14}\text{C}$  dating<sup>67</sup>. C:N ratios for well-preserved samples fall between 2.9 and 3.6, indicating good collagen preservation<sup>68</sup>. For 119 of the new dates, we also report  $\delta^{13}\text{C}$  and  $\delta^{15}\text{N}$  values (Supplementary Table 6).

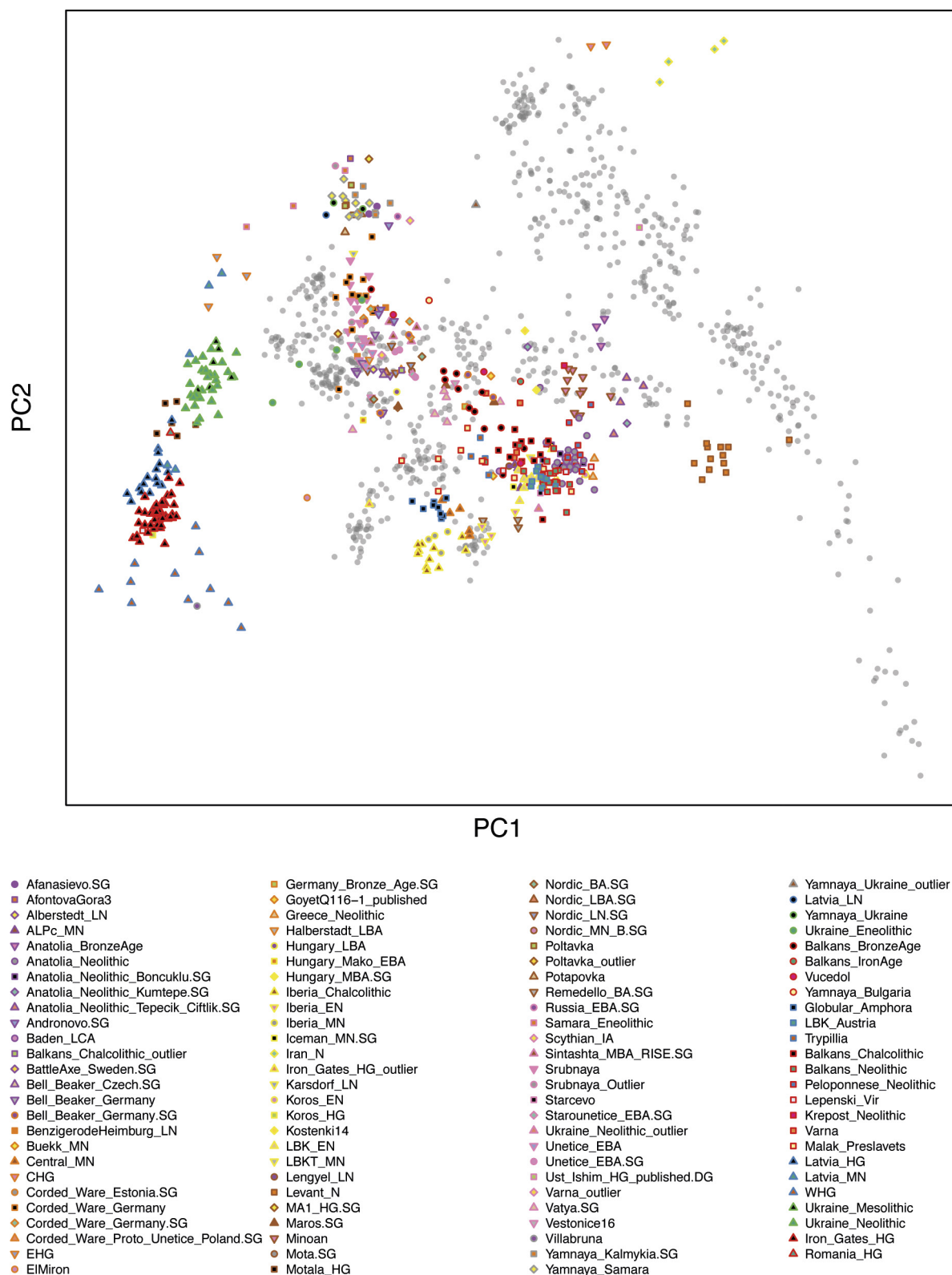
All  $^{14}\text{C}$  ages were  $\delta^{13}\text{C}$ -corrected for mass-dependent fractionation with measured  $^{13}\text{C}/^{12}\text{C}$  values<sup>69</sup> and calibrated with OxCal version 4.2.3<sup>70</sup> using the IntCal13 northern hemisphere calibration curve<sup>70</sup>. For hunter-gatherers from the Iron Gates, the direct  $^{14}\text{C}$  dates tend to be overestimates because of the freshwater reservoir effect (FRE), which arises because of a diet including fish that consumed ancient carbon. For these individuals, we performed a correction<sup>71</sup> (Supplementary Note 1) that assumed that 100% FRE = 545 ± 70 years, and  $\delta^{15}\text{N}$  values of 8.3‰ and 17.0‰ for 100% terrestrial and aquatic diets, respectively.

**Code availability.** The software used to analyse the data is available from the following sources: smartpca, qpAdm, qpDstat and qpGraph (<https://github.com/DRreichLab/AdmixTools/>), ADMIXTURE (<https://www.genetics.ucla.edu/software/admixture/>), EEMS (<https://github.com/dipetkov/eems/>), bwa (<http://bio-bwa.sourceforge.net>) and OxCal (<https://c14.arch.ox.ac.uk/oxcal.html>).

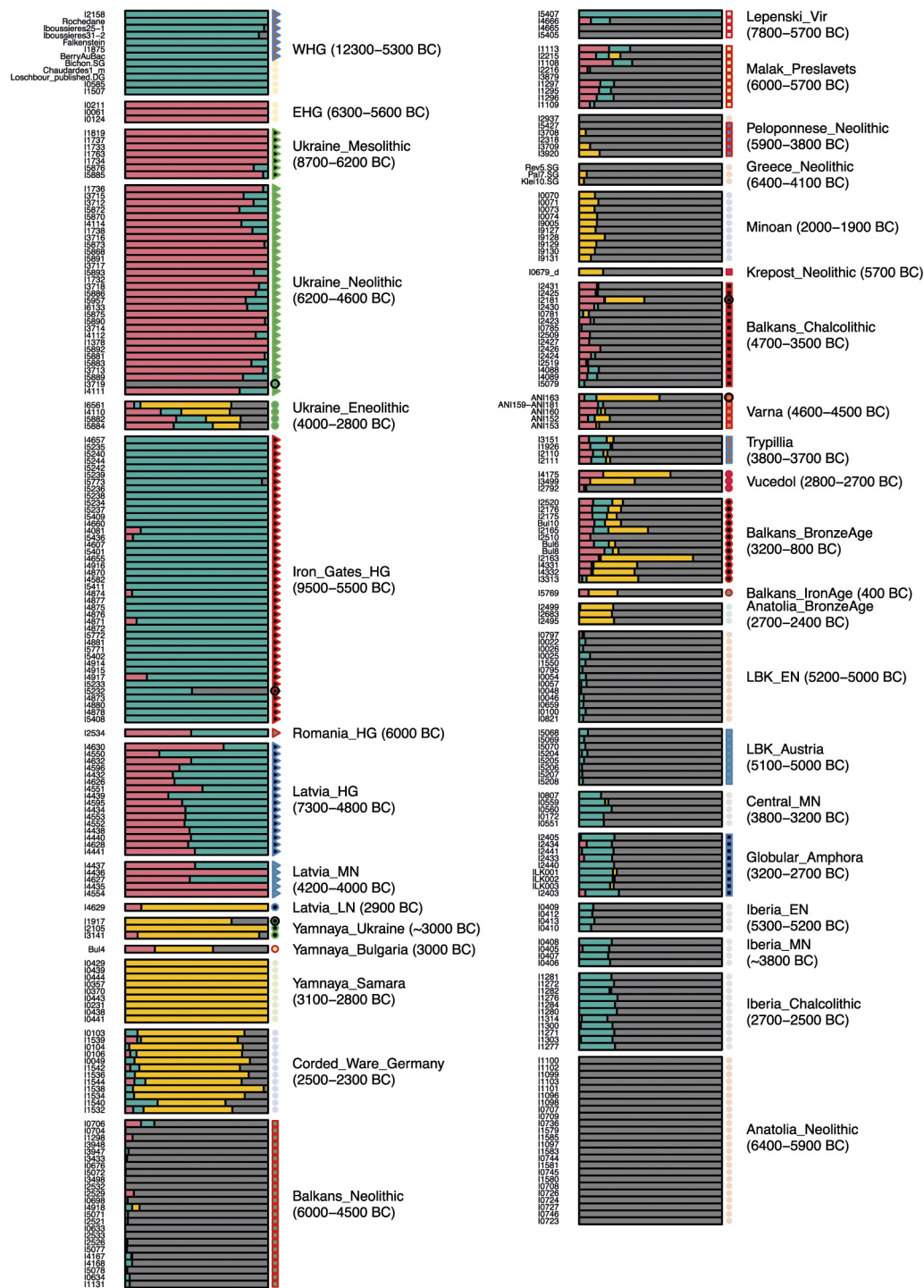
**Data availability.** The aligned sequences are available through the European Nucleotide Archive under accession number PRJEB22652. The pseudo-haploid genotype dataset used in analysis and in consensus mitochondrial genomes is available at <https://reich.hms.harvard.edu/datasets>.

48. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110**, 15758–15763 (2013).
49. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* <https://doi.org/10.1101/pdb.prot5448> (2010).
50. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Phil. Trans. R. Soc. Lond. B* **370**, 20130624 (2015).
51. Korlević, P. *et al.* Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *Biotechniques* **59**, 87–93 (2015).
52. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
53. DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* **23**, 4742–4743 (1995).
54. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
55. Maricic, T., Whitten, M. & Pääbo, S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* **5**, e14004 (2010).
56. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).
57. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
58. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013).
59. Skoglund, P. *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl Acad. Sci. USA* **111**, 2229–2234 (2014).
60. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
61. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
62. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
63. Brown, L. D., Cai, T. T. & DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **16**, 101–133 (2001).
64. Agresti, A. & Coull, B. A. Approximate is better than ‘exact’ for interval estimation of binomial proportions. *Am. Stat.* **52**, 119–126 (1998).
65. Longin, R. New method of collagen extraction for radiocarbon dating. *Nature* **230**, 241–242 (1971).
66. Brown, T. A., Nelson, D. E., Vogel, J. S. & Southon, J. R. Improved collagen extraction by modified Longin method. *Radiocarbon* **30**, 171–177 (1988).
67. Kennett, D. J. *et al.* Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nat. Commun.* **8**, 14115 (2017).
68. van Klinken, G. J. Bone collagen quality indicators for palaeodietary and radiocarbon measurements. *J. Archaeol. Sci.* **26**, 687–695 (1999).
69. Stuiver, M. & Polach, H. A. Discussion: reporting of  $^{14}\text{C}$  data. *Radiocarbon* **19**, 355–363 (1977).
70. Bronk Ramsey, C. *OxCal 4.23 Online Manual* [https://c14.arch.ox.ac.uk/oxcalhelp/hlp\\_contents.html](https://c14.arch.ox.ac.uk/oxcalhelp/hlp_contents.html) (2013).
71. Cook, G. T. *et al.* A freshwater diet-derived  $^{14}\text{C}$  reservoir effect at the Stone Age sites in the Iron Gates gorge. *Radiocarbon* **43**, 453–460 (2001).



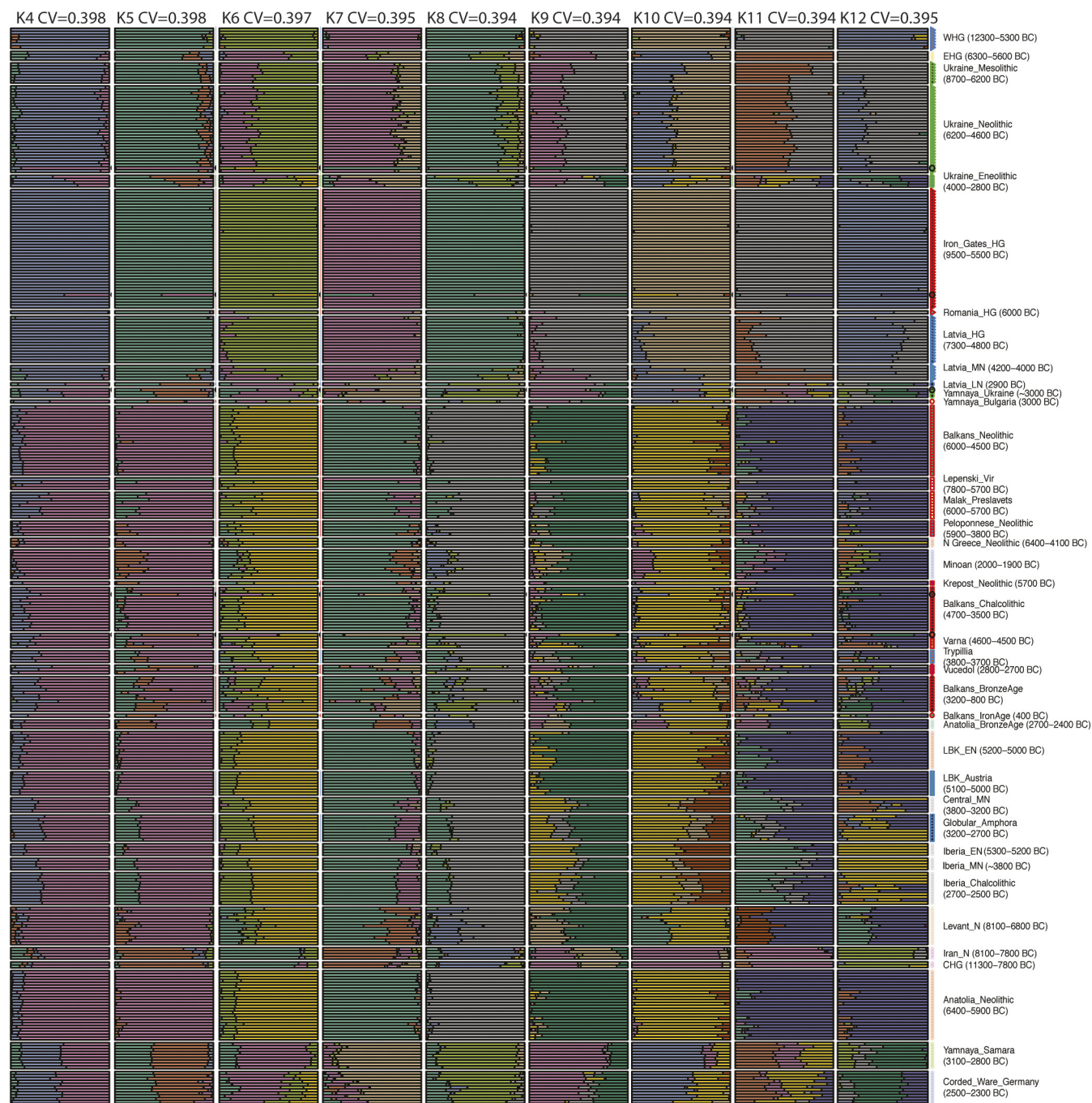


**Extended Data Figure 1 | Principal components analysis of ancient individuals.** Points for 486 ancient individuals are projected onto principal components defined by 777 present-day west Eurasian individuals (grey points). This differs from Fig. 1b in that the plot is not cropped and the present-day individuals are shown.



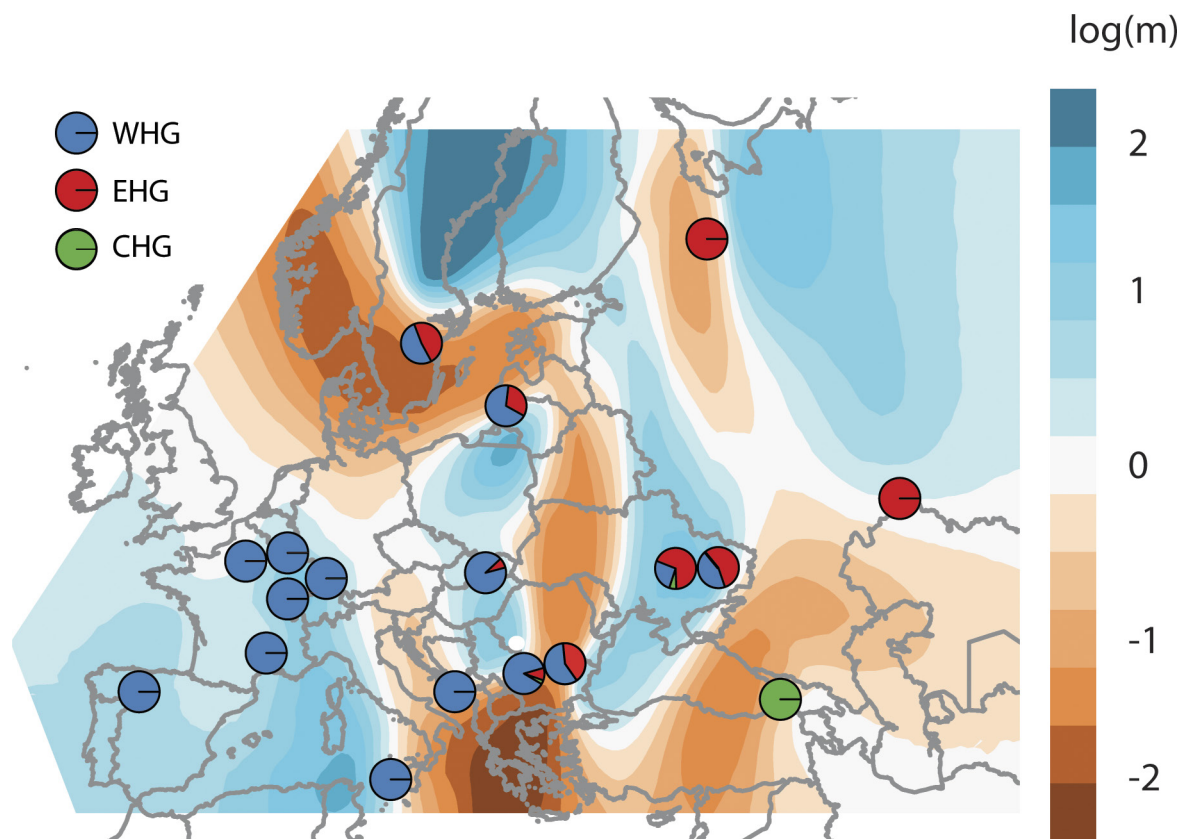
**Extended Data Figure 2 | Supervised ADMIXTURE analysis.** Supervised ADMIXTURE analysis modelling each ancient individual (one per row), as a mixture of populations represented by clusters that are constrained to contain northwestern-Anatolian Neolithic (grey), Yamnaya from

Samara (yellow), EHG (pink) and WHG (green) populations. Dates in parentheses indicate approximate range of individuals in each population. This differs from Fig. 1d in that it contains some previously published samples<sup>7,9,10,19,23,26</sup> and includes sample identification numbers.



**Extended Data Figure 3 | Unsupervised ADMIXTURE analysis.** Unsupervised ADMIXTURE plot from  $k = 4$  to 12 on a dataset consisting of 1,099 present-day individuals and 476 ancient individuals. We show newly reported ancient individuals and some previously published individuals<sup>7,10,19,22,23,26</sup> for comparison.

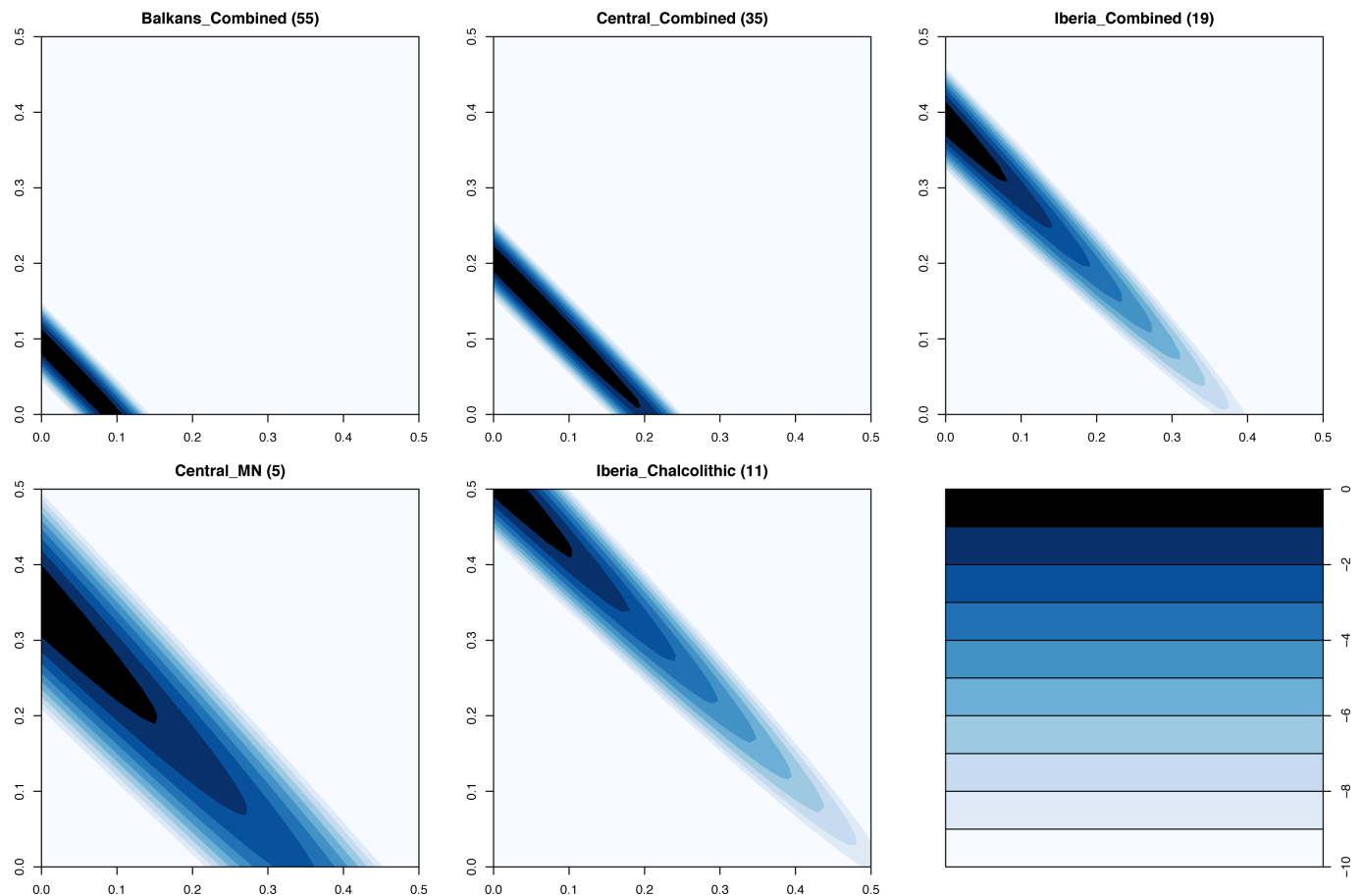




#### Extended Data Figure 4 | Genetic spatial structure in hunter-gatherers.

We infer the estimated effective migration surface<sup>62</sup>, a model of genetic relatedness in which individuals move in a random direction from generation to generation on an underlying grid, such that genetic relatedness is determined by distance. The migration parameter,  $m$ , defining the local rate of migration, varies on the grid and is inferred. This plot shows  $\log_{10}(m)$ , scaled relative to the average migration rate, which is arbitrary. Thus  $\log_{10}(m) = 2$ , for example, implies that the rate of migration at this point on the grid is 100 times higher than average. To restrict the model as much as possible to hunter-gatherer populations, the migration surface is inferred using data from 116 individuals that date to earlier than approximately 5000 BC and have no northwestern-Anatolian-Neolithic-related ancestry. Although the migration surface is

sensitive to sampling and fine-scale features may not be interpretable, the migration 'barrier' (region of low migration) running north-to-south and separating populations with primarily WHG ancestry from those with primarily EHG ancestry seems to be robust, and consistent with inferred admixture proportions. This analysis suggests that Mesolithic hunter-gatherer population structure was clustered and not smoothly clinal (that is, genetic differentiation did not vary consistently with distance). Superimposed on this background, pie charts show the WHG, EHG and CHG ancestry proportions inferred for populations used to construct the migration surface. This represents another way of visualizing the data in Fig. 2, Supplementary Table 3.1.3; we use two population models if they fit with  $P > 0.01$ , and three population models otherwise. Pie charts with only a single colour are those that were fixed to be the source populations.



**Extended Data Figure 5 | Sex bias in hunter-gatherer admixture.** The log-likelihood surfaces for the proportion of female ( $x$  axis) and male ( $y$  axis) ancestors that are hunter-gatherer-related for the combined populations analysed in Fig. 3c, and the two populations with the strongest

evidence for sex bias. Numbers in parentheses, number of individuals in each group. The log-likelihood scale ranges from 0 to  $-10$ , in which 0 is the feasible point with the highest likelihood.

# Hierarchical neural architecture underlying thirst regulation

Vineet Augustine<sup>1,2</sup>, Sertan Kutal Gokce<sup>2\*</sup>, Sangjun Lee<sup>2\*</sup>, Bo Wang<sup>2</sup>, Thomas J. Davidson<sup>3</sup>, Frank Reimann<sup>4</sup>, Fiona Gribble<sup>4</sup>, Karl Deisseroth<sup>5,6</sup>, Carlos Lois<sup>2</sup> & Yuki Oka<sup>1,2</sup>

Neural circuits for appetites are regulated by both homeostatic perturbations and ingestive behaviour. However, the circuit organization that integrates these internal and external stimuli is unclear. Here we show in mice that excitatory neural populations in the lamina terminalis form a hierarchical circuit architecture to regulate thirst. Among them, nitric oxide synthase-expressing neurons in the median preoptic nucleus (MnPO) are essential for the integration of signals from the thirst-driving neurons of the subfornical organ (SFO). Conversely, a distinct inhibitory circuit, involving MnPO GABAergic neurons that express glucagon-like peptide 1 receptor (GLP1R), is activated immediately upon drinking and monosynaptically inhibits SFO thirst neurons. These responses are induced by the ingestion of fluids but not solids, and are time-locked to the onset and offset of drinking. Furthermore, loss-of-function manipulations of GLP1R-expressing MnPO neurons lead to a polydipsic, overdrinking phenotype. These neurons therefore facilitate rapid satiety of thirst by monitoring real-time fluid ingestion. Our study reveals dynamic thirst circuits that integrate the homeostatic-instinctive requirement for fluids and the consequent drinking behaviour to maintain internal water balance.

The precise regulation of water intake is critical to the maintenance of fluid homeostasis in the body. The initiation of drinking in animals is triggered by internal fluid imbalance, such as water depletion<sup>1–4</sup>. By contrast, drinking is terminated rapidly when animals have ingested a sufficient amount of water, which generally precedes the absorption of the ingested fluid<sup>5–10</sup>. To achieve such accurate fluid regulation, the brain needs to monitor both internal water balance and fluid ingestion on a real-time basis<sup>11,12</sup>. How the brain integrates homeostatic and behavioural inputs to coordinate drinking behaviour is an unsolved question. As such, uncovering the neural circuits that process these regulatory signals is a critical step in understanding the neural logic of thirst regulation<sup>13–15</sup>.

The lamina terminalis is the principal brain structure responsible for sensing and regulating internal water balance<sup>3,5,16,17</sup>. It contains three main nuclei: the SFO, the organum vasculosum lamina terminalis (OVLT) and the MnPO, all of which are anatomically interconnected<sup>17–21</sup>. The SFO and the OVLT in particular are two major osmosensory sites in the brain because they lack the normal blood–brain barrier. Recent studies have shown that specific neural populations in the lamina terminalis have a causal role in the regulation of drinking behaviour. For instance, optogenetic and chemogenetic activation of excitatory SFO neurons co-expressing a transcription factor, ETV1, and nitric oxide synthase (SFO<sup>nNOS</sup> neurons) drives immediate and robust drinking behaviour<sup>19,22,23</sup>. Conversely, stimulation of inhibitory populations of lamina terminalis nuclei suppresses water intake<sup>19,24</sup>. Although these studies pinpointed the neural substrates that regulate thirst, the circuit organization that mediates drinking behaviour remains poorly understood, owing to anatomical complexity and the lack of genetic handles.

Here we focused on the neural architecture of the lamina terminalis, and investigated genetically defined thirst circuits using neural manipulation, tracing and *in vivo* optical recording approaches.

## Hierarchical circuit for thirst

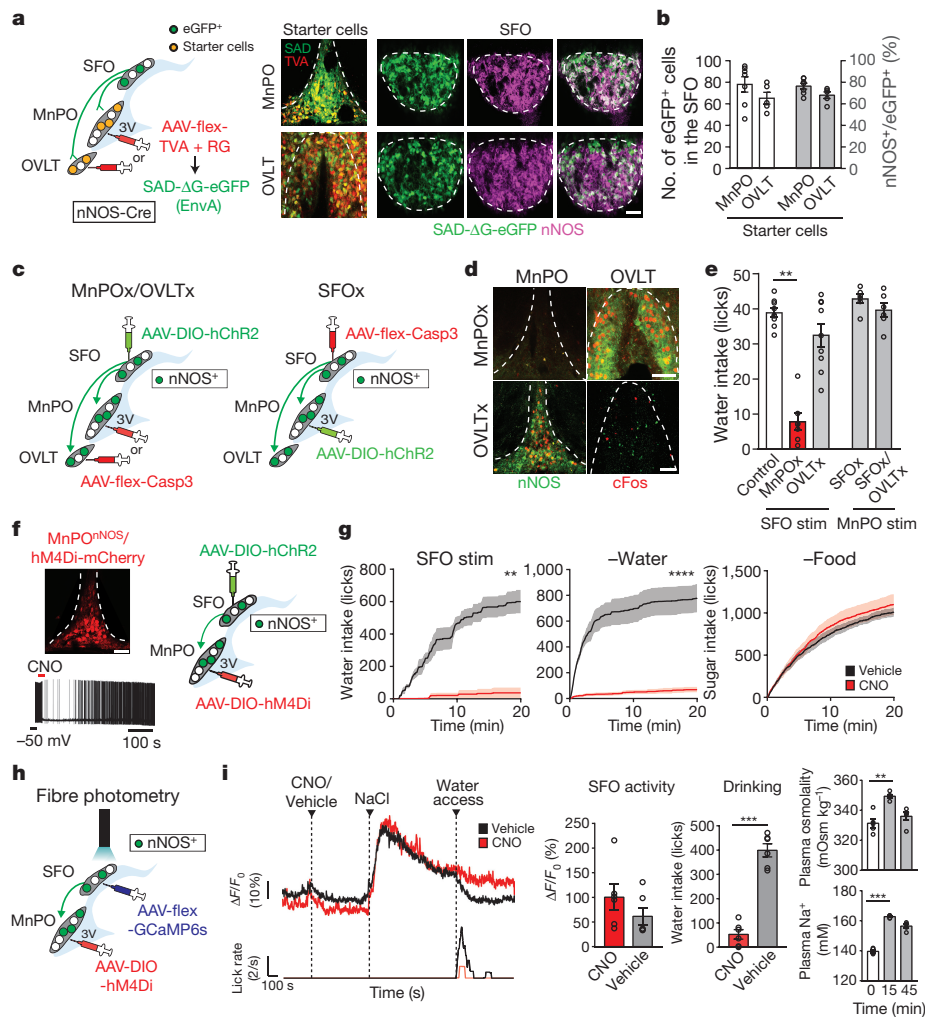
SFO<sup>nNOS</sup> neurons project their axons to other nuclei of the lamina terminalis (OVLT and MnPO)<sup>10,25</sup>, as well as to the paraventricular and supraoptic nuclei, which contain vasopressin-expressing neurons<sup>19</sup>. These axonal projections and the downstream neurons define a framework of circuit elements that control thirst-related behaviours and hormonal outputs<sup>26</sup>. To identify genetically defined SFO<sup>nNOS</sup> downstream populations that regulate drinking, we used optogenetics along with monosynaptic rabies tracing. Water restriction induces robust c-Fos expression in the SFO and putative downstream regions (Extended Data Fig. 1a). In the MnPO and OVLT, essentially all of the c-Fos signals were found in nNOS-expressing excitatory neurons (MnPO<sup>nNOS</sup> and OVLT<sup>nNOS</sup>; Extended Data Fig. 1a, top, b). Similar results were obtained when we photostimulated SFO<sup>nNOS</sup> neurons by expressing channel-rhodopsin (ChR2)<sup>27</sup> using adeno-associated virus (AAV-DIO-ChR2) in *nNOS-cre* (also known as *Nos1-cre*) mice (Extended Data Fig. 1a, bottom). These data suggest that MnPO<sup>nNOS</sup> and OVLT<sup>nNOS</sup> neurons are putative downstream populations of SFO<sup>nNOS</sup> neurons. Retrograde monosynaptic rabies tracing<sup>28</sup> from MnPO<sup>nNOS</sup> and OVLT<sup>nNOS</sup> neurons confirmed direct connections with the SFO<sup>nNOS</sup> population (Fig. 1a, b and Extended Data Fig. 1c). Moreover, photostimulation of ChR2-expressing MnPO<sup>nNOS</sup> or OVLT<sup>nNOS</sup> neurons selectively induced water drinking in satiated mice (Extended Data Fig. 1d). These studies demonstrated that SFO<sup>nNOS</sup> neurons send monosynaptic excitatory inputs to the MnPO<sup>nNOS</sup> and OVLT<sup>nNOS</sup> populations, each of which is sufficient to trigger water drinking.

To further investigate the circuit architecture that processes the internal need for water, we performed neural epistasis analysis for the circuits of the lamina terminalis by loss-of-function manipulation (Fig. 1c). We reasoned that if SFO<sup>nNOS</sup> and its downstream populations redundantly encode thirst in parallel, the ablation of one population should have only minor effects on drinking. Alternatively, if

<sup>1</sup>Computation and Neural Systems, California Institute of Technology, Pasadena, California, USA. <sup>2</sup>Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. <sup>3</sup>Department of Physiology and Kavli Institute for Fundamental Neuroscience, University of California, San Francisco, California, USA. <sup>4</sup>Department of Clinical Biochemistry, University of Cambridge, Cambridge, UK. <sup>5</sup>Howard Hughes Medical Institute, Stanford University, Stanford, California, USA. <sup>6</sup>Department of Bioengineering, Stanford University, Stanford, California, USA.

\*These authors contributed equally to this work.



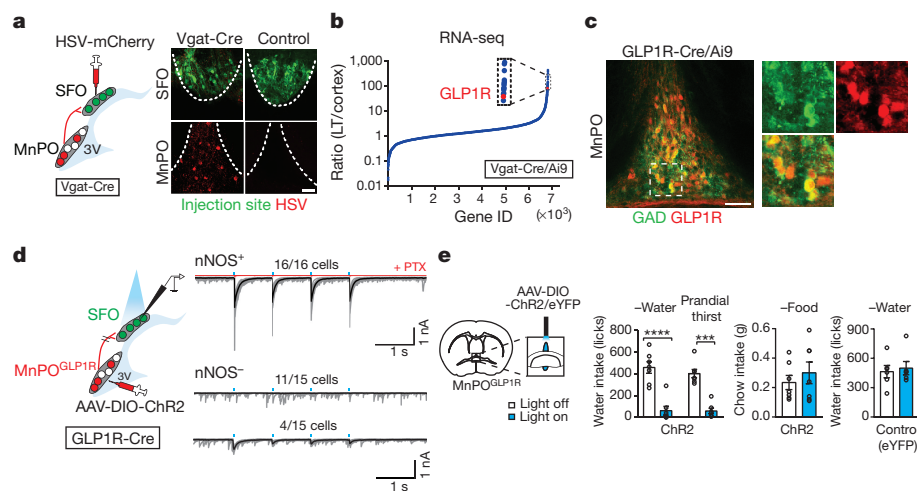


**Figure 1 | Thirst-driving neurons are organized hierarchically in the lamina terminalis.** **a**, Schematic of monosynaptic rabies tracing (left). Representative images of the MnPO (top right, one of seven mice) and OVLT (bottom right, one of five mice) of an *nNOS-cre* mouse transduced with AAV-CA-flex-RG and AAV-EF1a-flex-TVA-mCherry (red) followed by RV-SAD-ΔG-eGFP (green). 3V, third ventricle. **b**, Quantification of eGFP<sup>+</sup> neurons in the SFO ( $n = 7$  and  $5$  mice for MnPO and OVLT, respectively). **c**, Neural epistasis analysis of the circuits of the lamina terminalis by loss-of-function manipulation. Caspase expression is induced in the MnPO, OVLT (left) or SFO (right) of *nNOS-cre* mice. **d**, Casp3-TEVp efficiently eliminates nNOS-expressing neurons (green) in the MnPO ( $93.2 \pm 2.5\%$ ,  $n = 4$  mice) and OVLT ( $90.6 \pm 1.4\%$ ,  $n = 6$  mice). **e**, c-Fos expression (red) upon the stimulation of SFO<sup>nNOS</sup> neurons is shown. **f**, Number of licks during the 5-s session ( $n = 9$  mice for controls and OVLTx,  $n = 7$  mice for MnPOx,  $n = 6$  mice for SFOx and SFOx/OVLTx). **g**, Cumulative water intake in SFO<sup>nNOS</sup>-stimulated mice (left,  $n = 5$  mice) or water-restricted mice (middle,  $n = 10$  mice for CNO and  $n = 9$  mice for vehicle), and sucrose (300 mM) intake in food-restricted mice (right,  $n = 10$  mice for CNO and  $n = 9$  mice for vehicle). **h**, Fibre photometry of SFO<sup>nNOS</sup> neurons while MnPO<sup>nNOS</sup> neurons are inhibited by hM4Di-mCherry. **i**, Intraperitoneal NaCl injection robustly activates SFO<sup>nNOS</sup> neurons with (red trace) or without (black trace) CNO injection (left and middle left). By contrast, CNO injection drastically suppressed drinking behaviour (middle right,  $n = 6$  mice). Plasma osmolality (top right) and Na<sup>+</sup> concentration (bottom right) were measured after NaCl injection ( $n = 5$  mice). \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , \*\*\*\* $P < 0.0001$ , by Mann-Whitney  $U$  test, paired two-tailed  $t$ -test or Kruskal-Wallis one-way analysis of variance (ANOVA) test. All error bars and shaded areas show mean  $\pm$  s.e.m. Scale bars, 50  $\mu$ m.

the circuit is organized in a hierarchical fashion in which a specific population has a critical role, the elimination of such a downstream population should abolish SFO<sup>nNOS</sup>-stimulated drinking. To test these ideas, we expressed caspase (AAV-flex-Casp3)<sup>29</sup> in the MnPO, OVLT or SFO of *nNOS-cre* mice (Fig. 1c). The expression of Casp3 resulted in the specific and near-complete elimination of nNOS-expressing neurons of a given nucleus (Fig. 1d and Extended Data Fig. 2a). In OVLT<sup>nNOS</sup>-ablated and control mice, photostimulation of SFO<sup>nNOS</sup> neurons triggered robust drinking (Fig. 1e and Extended Data Fig. 2b). By sharp contrast, the ablation of MnPO<sup>nNOS</sup> neurons markedly suppressed SFO<sup>nNOS</sup>-stimulated water intake (Fig. 1e and Extended Data Fig. 2b, MnPOx). We also found that MnPO<sup>nNOS</sup> neurons have an important role in the drinking behaviour evoked by OVLT<sup>nNOS</sup> neurons. Water intake induced by photostimulation of OVLT<sup>nNOS</sup>

neurons was significantly attenuated after ablating MnPO<sup>nNOS</sup>, but not SFO<sup>nNOS</sup> neurons (Extended Data Fig. 2c). These results suggest that MnPO<sup>nNOS</sup> neurons are essential neural substrates of the lamina terminalis for the behavioural output. If this model is correct, stimulating the MnPO<sup>nNOS</sup> population without the inputs from their upstream SFO<sup>nNOS</sup>, or both SFO<sup>nNOS</sup> and OVLT<sup>nNOS</sup> neurons should still trigger robust drinking (Fig. 1c). As hypothesized, the elimination of these populations had no impact on drinking when MnPO<sup>nNOS</sup> neurons were directly photostimulated (Fig. 1e and Extended Data Fig. 2b, SFOx, SFOx and OVLTx). Similar results were obtained by chemogenetic acute silencing using hM4Di (ref. 30) (Fig. 1f). In awake mice, acute inhibition of MnPO<sup>nNOS</sup> neurons by clozapine *N*-oxide (CNO) severely suppressed water consumption in both water-restricted and SFO<sup>nNOS</sup>-stimulated mice (Fig. 1g and Extended Data Fig. 2d, e). However, the

out of six neurons), and a diagram of photostimulation of SFO<sup>nNOS</sup> and chemogenetic inhibition of MnPO<sup>nNOS</sup> neurons (right). **g**, Cumulative water intake in SFO<sup>nNOS</sup>-stimulated mice (left,  $n = 5$  mice) or water-restricted mice (middle,  $n = 10$  mice for CNO and  $n = 9$  mice for vehicle), and sucrose (300 mM) intake in food-restricted mice (right,  $n = 10$  mice for CNO and  $n = 9$  mice for vehicle). **h**, Fibre photometry of SFO<sup>nNOS</sup> neurons while MnPO<sup>nNOS</sup> neurons are inhibited by hM4Di-mCherry. **i**, Intraperitoneal NaCl injection robustly activates SFO<sup>nNOS</sup> neurons with (red trace) or without (black trace) CNO injection (left and middle left). By contrast, CNO injection drastically suppressed drinking behaviour (middle right,  $n = 6$  mice). Plasma osmolality (top right) and Na<sup>+</sup> concentration (bottom right) were measured after NaCl injection ( $n = 5$  mice). \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , \*\*\*\* $P < 0.0001$ , by Mann-Whitney  $U$  test, paired two-tailed  $t$ -test or Kruskal-Wallis one-way analysis of variance (ANOVA) test. All error bars and shaded areas show mean  $\pm$  s.e.m. Scale bars, 50  $\mu$ m.



**Figure 2 | GLP1R-expressing GABAergic neurons in the MnPO are a major source of inhibitory input to the SFO.** **a**, GABAergic input to the SFO. Representative image of the SFO and MnPO after co-injection of AAV-Syn-GCaMP6s (green) and HSV-mCherry (red) in the SFO (one out of four mice). **b**, GLP1R expression is enriched in inhibitory neurons from the lamina terminalis (LT) relative to the cortex. **c**, MnPO<sup>GLP1R</sup> neurons are GABAergic (84.7 ± 3.4% of GAD<sup>+</sup> neurons are tdTomato<sup>+</sup>, *n* = 3 mice, representative images are from one out of three mice). These neurons

same manipulation did not decrease sugar consumption in food-restricted mice (Fig. 1g and Extended Data Fig. 2d, e).

Importantly, the silencing of MnPO<sup>nNOS</sup> neurons did not compromise the osmosensory function of the SFO<sup>nNOS</sup> population. We used fibre photometry<sup>31</sup> in awake-behaving mice that expressed the calcium indicator GCaMP6s in the SFO<sup>nNOS</sup>, and the neuronal silencer hM4Di in MnPO<sup>nNOS</sup> neurons (Fig. 1h). We showed that the activation of SFO<sup>nNOS</sup> neurons by osmotic stress was unaffected in the absence of functioning MnPO<sup>nNOS</sup> neurons (Fig. 1i and Extended Data Fig. 2f). These results were supported by our electrophysiological recordings: only a minor fraction of SFO neurons received monosynaptic input from MnPO<sup>nNOS</sup> neurons (Extended Data Fig. 3), demonstrating the unidirectional connection from SFO<sup>nNOS</sup> to MnPO<sup>nNOS</sup> neurons. Taken together, our results demonstrate that thirsty neurons in the lamina terminalis form a hierarchical circuit organization, and that the MnPO<sup>nNOS</sup> population is required to process signals from SFO<sup>nNOS</sup> neurons to coordinate drinking.

### MnPO<sup>GLP1R</sup> → SFO<sup>nNOS</sup> inhibitory input

The thirst neurons of the lamina terminalis also receive negative feedback regulation upon drinking itself<sup>1,8,10</sup>. It has been shown that water intake rapidly suppresses the activity of thirst neurons in the lamina terminalis<sup>10,18</sup> (Extended Data Fig. 4). It is suggested that this quick regulation of thirst circuits optimizes fluid ingestion<sup>8,9</sup>. To examine the neural basis of drinking-induced thirst inhibition, we functionally mapped the upstream inhibitory circuits of SFO<sup>nNOS</sup> neurons using two neural tracing approaches. First, we retrogradely labelled inhibitory neurons that project to the SFO by injecting herpes simplex virus conjugated with mCherry (HSV-mCherry) into the SFO of *Vgat-cre* mice (Fig. 2a, left). Among the putative upstream structures (Extended Data Fig. 5a), the MnPO contained the strongest HSV signals (Fig. 2a, right). Next, we performed monosynaptic rabies tracing from SFO<sup>nNOS</sup> neurons (Extended Data Fig. 5b). Consistent with the results of the HSV tracing, the MnPO contained the greatest number of rabies-virus-positive neurons that minimally overlapped with excitatory neurons (Extended Data Fig. 5b). These complementary tracing results suggest that GABAergic neurons in the MnPO are a major source of inhibitory input to the SFO<sup>24</sup>.

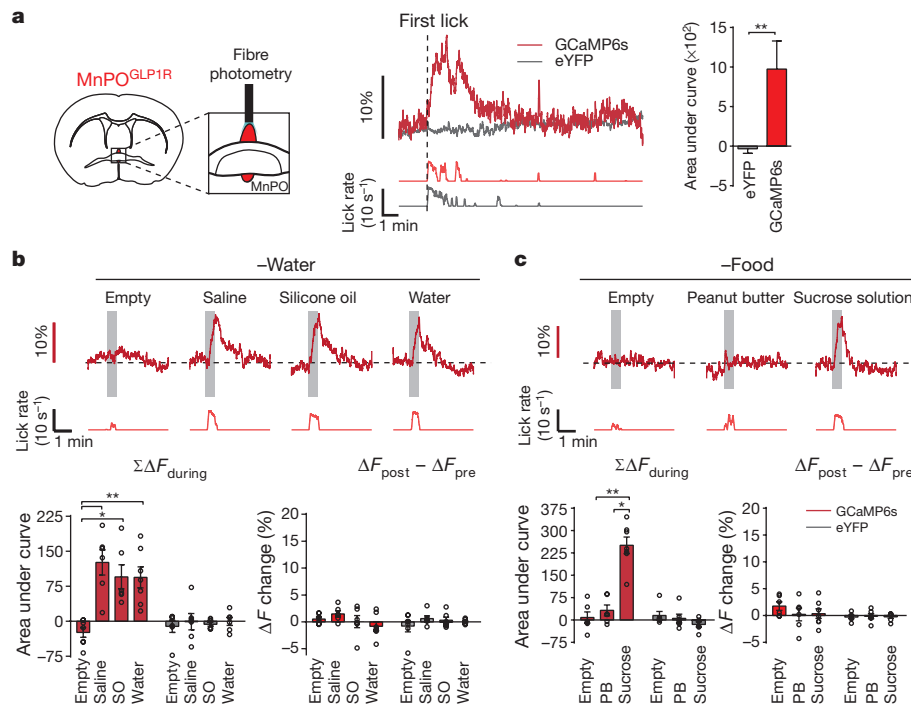
To gain a more specific genetic handle on these neurons, we performed RNA sequencing analysis of the inhibitory population of the dorsal lamina terminalis (containing the MnPO and SFO) and the

did not overlap with glutamatergic neurons (4.3 ± 0.9% overlap, *n* = 3 mice, Extended Data Fig. 6a). **d**, The MnPO<sup>GLP1R</sup> → SFO monosynaptic connection. MnPO<sup>GLP1R</sup> neurons send monosynaptic inhibitory input to SFO<sup>nNOS</sup> neurons. **e**, Optogenetic stimulation of MnPO<sup>GLP1R</sup> neurons selectively suppresses water intake (*n* = 7 mice for ChR2 and *n* = 6 mice for control). \*\*\**P* < 0.001, \*\*\*\**P* < 0.0001, by paired two-tailed *t*-test. All error bars show mean ± s.e.m. Scale bars, 50 μm.

cortex. We found that GLP1R transcripts were highly enriched in the inhibitory neurons from the lamina terminalis, by a factor of 100 compared to the cortex (Fig. 2b). *In situ* hybridization and immunohistochemical studies in *Glp1r-cre* mice<sup>32</sup> confirmed that GABAergic MnPO neurons expressed GLP1R (Fig. 2c and Extended Data Fig. 6a, b). As predicted from our tracing results, ChR2-assisted circuit mapping<sup>33</sup> revealed that all recorded SFO<sup>nNOS</sup> neurons (16 out of 16 cells) received robust monosynaptic inhibitory input from GLP1R-expressing MnPO (MnPO<sup>GLP1R</sup>) neurons, with an inhibitory postsynaptic current latency of 8.4 ms (Fig. 2d). However, SFO<sup>non-nNOS</sup> neurons received such input rarely (4 out of 15 cells with small inhibitory postsynaptic currents, Fig. 2d), showing that inhibitory input from MnPO<sup>GLP1R</sup> neurons is specific to excitatory neurons in the SFO. Furthermore, photostimulation of MnPO<sup>GLP1R</sup> neurons selectively suppressed water intake in thirsty mice (Fig. 2e and Extended Data Fig. 6c), although this acute inhibition was not observed upon the application of a GLP1R agonist<sup>34</sup> (Extended Data Fig. 6d–f). Collectively, our findings suggest that the MnPO<sup>GLP1R</sup> population has a key modulatory role in thirst.

### MnPO<sup>GLP1R</sup> neurons monitor liquid intake

Next, we measured the *in vivo* calcium dynamics of MnPO<sup>GLP1R</sup> neurons expressing GCaMP6s in *Glp1r-cre* mice (Fig. 3a). In freely moving mice, MnPO<sup>GLP1R</sup> neurons were acutely activated during water drinking, and their activity returned to the basal level when they stopped drinking (Fig. 3a, red trace). These neurons responded equally when thirsty mice licked either water or isotonic saline, but not when they licked an empty spout (Fig. 3b and Extended Data Fig. 7c–e). Notably, the neuronal responses were also observed when the mice licked non-aqueous silicone oil, which showed that the activation of MnPO<sup>GLP1R</sup> neurons is independent of fluid composition. Under food-restricted conditions, we found that MnPO<sup>GLP1R</sup> neurons still responded upon licking sucrose solution (300 mM, Fig. 3c and Extended Data Fig. 7c, d). However, solid peanut butter evoked no response despite its high palatability (Fig. 3c). These optical recording studies indicate that MnPO<sup>GLP1R</sup> neurons are activated purely by fluid consumption and not by reward-seeking behaviour or licking action per se. Consistent with the connection from MnPO<sup>GLP1R</sup> to SFO<sup>nNOS</sup> neurons, the activity of the SFO<sup>nNOS</sup> population mirrored precisely the calcium dynamics of MnPO<sup>GLP1R</sup> neurons, except that water intake evoked an additional persistent inhibition (Extended Data Fig. 7a, b).



**Figure 3 | Rapid and transient activation of MnPO<sup>GLP1R</sup> neurons during drinking behaviour.** **a**, Fibre photometry recording from MnPO<sup>GLP1R</sup> neurons (left). MnPO<sup>GLP1R</sup> neurons are activated upon drinking behaviour (right). Representative traces are from GCaMP6s and enhanced yellow fluorescent protein (eYFP) (one out of six mice). **b**, Responses of MnPO<sup>GLP1R</sup> neurons under water-restricted conditions towards different types of liquid. Transient activation (bottom left,  $\Sigma\Delta F_{\text{during}}$ ) and baseline activity shift (bottom right,  $\Delta F_{\text{post}} - \Delta F_{\text{pre}}$ ) were quantified ( $n = 6$  mice

for saline and silicone oil (SO),  $n = 7$  mice for empty and water,  $n = 6$  mice for all eYFP controls). **c**, Representative responses of MnPO<sup>GLP1R</sup> neurons under food-restricted conditions. Transient activation (bottom left,  $\Sigma\Delta F_{\text{during}}$ ) and baseline activity shift (bottom right,  $\Delta F_{\text{post}} - \Delta F_{\text{pre}}$ ) were quantified ( $n = 6$  mice for empty and peanut butter (PB),  $n = 7$  mice for sucrose,  $n = 6$  mice for all eYFP controls). \* $P < 0.05$ , \*\* $P < 0.01$ , by two-tailed Mann–Whitney  $U$  test or Kruskal–Wallis one-way ANOVA test. All error bars show mean  $\pm$  s.e.m.

This water-specific inhibition of SFO<sup>nNOS</sup> neurons is probably due to osmolality sensing or water absorption in the gastrointestinal tract as proposed previously<sup>1,9</sup>. These results demonstrate two important properties of thirst circuits. First, MnPO<sup>GLP1R</sup> neurons are activated upon fluid ingestion; this activation is independent of fluid composition and the internal state of the animal. Second, this neural population transmits inhibitory signals to SFO<sup>nNOS</sup> neurons, in a manner that is time-locked to drinking.

### Effect of eating and drinking

We investigated the mechanisms by which MnPO<sup>GLP1R</sup> neurons exclusively represent fluid intake. To this end, we provided water-restricted mice with water in two different forms—liquid and gel (HydroGel: 98% water + hydrocolloids)—while recording MnPO<sup>GLP1R</sup> activity (Fig. 4a). In either form, the mice ingested a similar amount of water within the 30-min session (Fig. 4b). Notably, compared to the robust activation of MnPO<sup>GLP1R</sup> neurons upon drinking water, gel-eating behaviour did not elicit any response (Fig. 4a, c). Similarly, eating normal chow did not stimulate this neural population (Fig. 4d). Therefore, MnPO<sup>GLP1R</sup> neurons are able to distinguish between drinking and eating behaviour even if an animal consumes essentially the same substance. These results suggest that the MnPO<sup>GLP1R</sup> population facilitates satiety, which is induced by drinking behavior and not specifically by water.

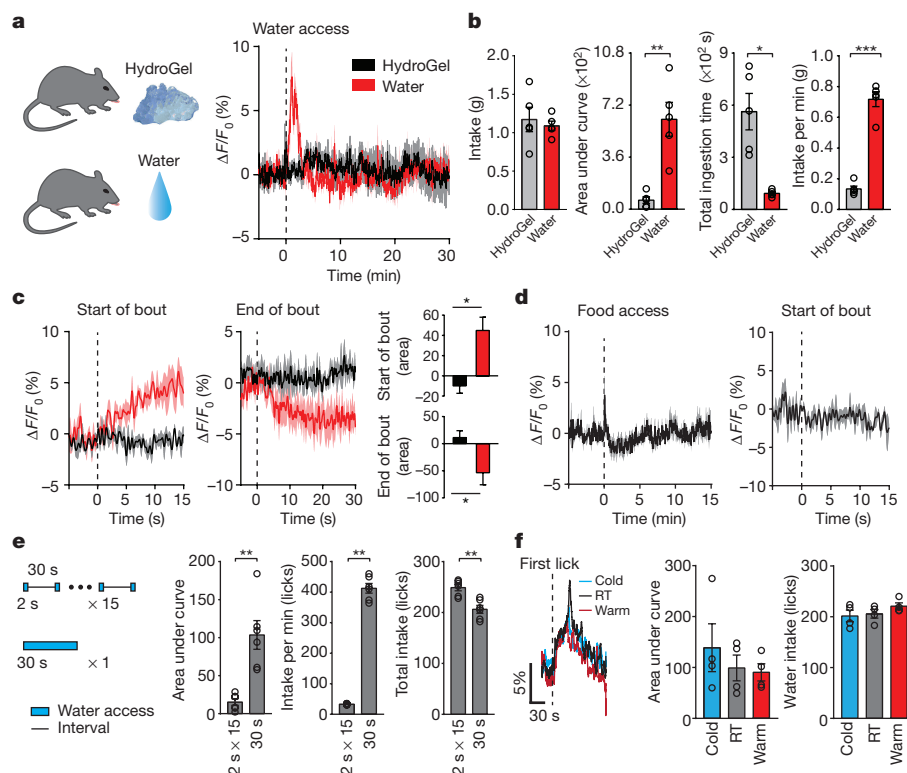
Because the rate of ingestion differed considerably between the drinking of water and the eating of HydroGel (Fig. 4b), we speculated that MnPO<sup>GLP1R</sup> neurons may monitor the pattern of ingestion in order to distinguish the mode of consumption. To examine this possibility, mice were given access to water for 30 s in total at two different rates:  $2 \times 15$  times and  $30 \times 1$  time (Fig. 4e). As hypothesized, concentrated periods of drinking evoked significantly greater responses in the MnPO<sup>GLP1R</sup> neurons than did sparse periods of drinking, regardless of

the total amount of water consumed (Fig. 4e). We note that the temperature of the fluid did not affect the response (Fig. 4f). Because animals can ingest fluids much faster than they can ingest solid substances, these data strongly support the idea that the MnPO<sup>GLP1R</sup> population distinguishes between drinking and eating on the basis of ingestion speed. Consequently, concentrated (that is, rapid) fluid intake recruits MnPO<sup>GLP1R</sup>-mediated inhibition signals, which in turn suppress the activity of SFO<sup>nNOS</sup> neurons. These findings provide key mechanistic insight into rapid thirst alleviation as a result of drinking behaviour.

### MnPO<sup>GLP1R</sup> neurons help thirst satiety

In view of the function of the MnPO<sup>GLP1R</sup> population in the monitoring of fluid intake, we next considered its physiological importance in the regulation of drinking using chemogenetic loss-of-function manipulation (Fig. 5a). Whereas any fluid elicits transient MnPO<sup>GLP1R</sup>  $\rightarrow$  SFO<sup>nNOS</sup> inhibition, water evokes an additional inhibitory effect that persists after drinking episodes (Extended Data Fig. 7a). Owing to this water-specific signal, inhibition of MnPO<sup>GLP1R</sup> neurons by CNO had only a minor effect on the total water intake of water-restricted mice during a 30-minute period (Extended Data Fig. 8a, b, d). By contrast, marked effects were observed for isotonic saline, in which MnPO<sup>GLP1R</sup>-independent inhibitory signals are absent (Fig. 5b). Compared to the vehicle control, inhibition of MnPO<sup>GLP1R</sup> neurons robustly increased both the total amount and the duration of saline intake (Fig. 5c and Extended Data Fig. 8c). However, under satiated conditions, the same manipulation did not increase water or saline intake, which excludes the possibility that inhibiting MnPO<sup>GLP1R</sup> neurons stimulates appetite directly (Fig. 5c). We observed the same overdrinking phenotype in mice in which MnPO<sup>GLP1R</sup> neurons were ablated by Casp3 (Extended Data Fig. 8e, f). Our functional manipulation studies demonstrate that MnPO<sup>GLP1R</sup> neurons promote satiety of thirst by monitoring real-time fluid intake, and that the malfunction of this neuronal regulation leads





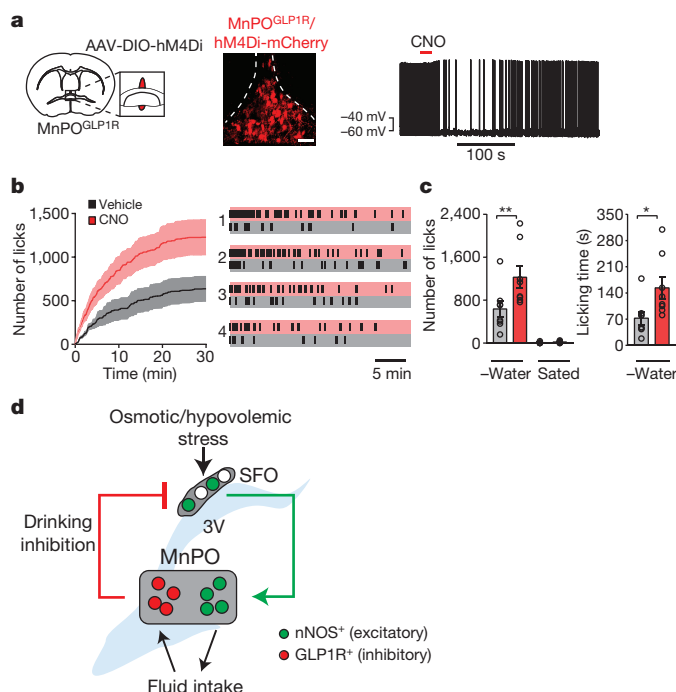
**Figure 4 | MnPO<sup>GLP1R</sup> neurons distinguish between drinking and eating behaviour based on ingestive speed.** **a**, MnPO<sup>GLP1R</sup> neurons respond to the intake of liquid water (red) but not HydroGel (black). **b**, Quantification of neural activity and drinking behaviour for the ingestion of HydroGel or water ( $n = 5$  mice). **c**, Peristimulus time histogram around the start (left) and the end (middle) of water and gel intake ( $n = 5$  mice); quantified data are shown on the right. **d**, Eating solid chow does not stimulate MnPO<sup>GLP1R</sup> neurons ( $n = 5$  mice). **e**, MnPO<sup>GLP1R</sup> neurons are stimulated to a greater extent during periods of concentrated drinking compared with sparse drinking ( $n = 6$  mice). **f**, The temperature of the ingested fluid has no effect on MnPO<sup>GLP1R</sup> activity. Total responses (middle) and the number of licks (right) were quantified ( $n = 4$  mice). \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , by two-tailed Mann–Whitney  $U$  test or paired two-tailed  $t$ -test. All error bars and shaded areas show mean  $\pm$  s.e.m.

to polydipsic overdrinking, especially in the case of non-hypoosmotic fluids such as saline.

## Discussion

In this study, we identified genetically defined thirst circuits in the lamina terminalis that integrate the instinctive need for water with the consequent drinking behaviour to maintain internal water balance (Fig. 5d). We showed that multiple downstream populations of SFO<sup>nNOS</sup> neurons are individually sufficient to induce water intake. These data are reminiscent of the circuit organization for hunger, in which eating behaviour is redundantly encoded by multiple output projections of AgRP neurons in the arcuate nucleus<sup>35</sup>. However, we showed that individual thirst-related neuronal populations of the lamina terminalis are hierarchically organized, and that MnPO<sup>nNOS</sup> neurons are the behavioural output neurons. Previous lesion studies in rats and sheep have proposed a model in which the MnPO serves as a critical site that integrates inputs from osmosensory neurons of the SFO and the OVLT<sup>36–38</sup>. Our findings well explain and further advance the concept of this model with cell-type-specific precision. Whereas the necessity of the SFO may vary among species<sup>10</sup>, the MnPO appears to consistently function as the key centre for drinking across species<sup>38</sup>. In our analysis, MnPO<sup>nNOS</sup> neurons project to various areas including the hypothalamus and the midbrain (Extended Data Fig. 9a; see also ref. 18). These results reveal a neural logic to thirst processing in the lamina terminalis circuit, and provide a platform for investigation into how the appetite for water is integrated at downstream sites of MnPO<sup>nNOS</sup> neurons.

Notably, MnPO<sup>GLP1R</sup> neurons responded selectively to the ingestion of fluids but not solids. These inhibitory neurons provide rapid monosynaptic inhibition to thirst-driving SFO<sup>nNOS</sup> neurons. Our results indicate strongly that the MnPO<sup>GLP1R</sup> population facilitates thirst satiation upon drinking rather than upon water absorption. At a psychophysical level, these findings provide an explanation for the long-standing observation that thirst is quickly alleviated at the onset of drinking<sup>6,9</sup>. At a physiological level, these results reveal a neural interface that adjusts the activity of thirst neurons on the basis of real-time drinking behaviour. Although systemic recovery of fluid balance relies on water absorption into the blood, thirst is modulated by multiple preabsorptive factors including oral, oropharyngeal and gastrointestinal signals<sup>1</sup>. It is unlikely



**Figure 5 | Inhibition of MnPO<sup>GLP1R</sup> neurons leads to overdrinking.** **a**, Treatment with CNO inhibits firing in hM4Di-expressing MnPO<sup>GLP1R</sup> neurons (right, 6 out of 7 neurons). **b**, Acute inhibition of MnPO<sup>GLP1R</sup> neurons by CNO results in the overdrinking of isotonic saline in water-restricted mice ( $n = 8$  mice). Representative lick patterns from four out of eight mice are shown (right). **c**, The total amount of saline intake and the time spent drinking ( $n = 8$  mice). **d**, A schematic summarizing thirst genesis, detection of fluid intake and drinking-induced feedback inhibition in the lamina terminalis circuit. \* $P < 0.05$ , \*\* $P < 0.01$ , by paired two-tailed  $t$ -test. All error bars and shaded areas show mean  $\pm$  s.e.m. Scale bar, 50  $\mu$ m.

that the  $\text{MnPO}^{\text{GLPIR}} \rightarrow \text{SFO}^{\text{nNOS}}$  circuit mediates oral sensory information such as taste<sup>39–41</sup> because it responds to any fluid, including silicone oil. Instead,  $\text{MnPO}^{\text{GLPIR}}$  neurons may function as a flow-meter by sensing gulping actions in the oropharyngeal area, and provide rapid, liquid-specific inhibition to thirst circuits. This idea is consistent with previous findings that drinking hyperosmotic saline<sup>7</sup>, but not eating food<sup>42</sup>, transiently suppressed vasopressin secretion. In this model,  $\text{MnPO}^{\text{GLPIR}}$  neurons serve as a central detector that discriminates fluid ingestion from solid ingestion, which promotes acute satiation of thirst through the SFO and other downstream targets (Extended Data Fig. 9b). Subsequently, gastrointestinal mechanisms may selectively detect water over other fluids that induce persistent inhibitory effects on  $\text{SFO}^{\text{nNOS}}$  neurons (Extended Data Fig. 7a). Although fluid-sensing mechanisms at each peripheral area are poorly understood, further molecular and cellular studies should help to reveal complex regulatory signals that maintain body-fluid homeostasis.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 3 August 2017; accepted 3 January 2018.**

**Published online 28 February 2018.**

- Ramsay, D. J. & Booth, D. (eds) *Thirst: Physiological and Psychological Aspects*. Ch. 5, 6, 9–12, 19 (Springer, 1991).
- Bourque, C. W. Central mechanisms of osmosensation and systemic osmoregulation. *Nat. Rev. Neurosci.* **9**, 519–531 (2008).
- Fitzsimons, J. T. Angiotensin, thirst, and sodium appetite. *Physiol. Rev.* **78**, 583–686 (1998).
- McKinley, M. J. & Johnson, A. K. The physiological regulation of thirst and fluid intake. *News Physiol. Sci.* **19**, 1–6 (2004).
- Johnson, A. K. & Gross, P. M. Sensory circumventricular organs and brain homeostatic pathways. *FASEB J.* **7**, 678–686 (1993).
- Saker, P. et al. Regional brain responses associated with drinking water during thirst and after its satiation. *Proc. Natl Acad. Sci. USA* **111**, 5379–5384 (2014).
- Seckl, J. R., Williams, T. D. & Lightman, S. L. Oral hypertonic saline causes transient fall of vasopressin in humans. *Am. J. Physiol.* **251**, R214–R217 (1986).
- Stricker, E. M. & Hoffmann, M. L. Presystemic signals in the control of thirst, salt appetite, and vasopressin secretion. *Physiol. Behav.* **91**, 404–412 (2007).
- Thrasher, T. N., Nistal-Herrera, J. F., Keil, L. C. & Ramsay, D. J. Satiety and inhibition of vasopressin secretion after drinking in dehydrated dogs. *Am. J. Physiol.* **240**, E394–E401 (1981).
- Zimmerman, C. A. et al. Thirst neurons anticipate the homeostatic consequences of eating and drinking. *Nature* **537**, 680–684 (2016).
- Farrell, M. J. et al. Cortical activation and lamina terminalis functional connectivity during thirst and drinking in humans. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **301**, R623–R631 (2011).
- Gizowski, C. & Bourque, C. W. The neural basis of homeostatic and anticipatory thirst. *Nat. Rev. Nephrol.* **14**, 11–25 (2018).
- Andermann, M. L. & Lowell, B. B. Toward a wiring diagram understanding of appetite control. *Neuron* **95**, 757–778 (2017).
- Sternson, S. M. Hypothalamic survival circuits: blueprints for purposive behaviors. *Neuron* **77**, 810–824 (2013).
- Zimmerman, C. A., Leib, D. E. & Knight, Z. A. Neural circuits underlying thirst and fluid homeostasis. *Nat. Rev. Neurosci.* **18**, 459–469 (2017).
- Denton, D. A., McKinley, M. J. & Weisinger, R. S. Hypothalamic integration of body fluid regulation. *Proc. Natl Acad. Sci. USA* **93**, 7397–7404 (1996).
- McKinley, M. J. et al. in *The Sensory Circumventricular Organs of the Mammalian Brain* Vol. 172 (ed. McKinley, M. J.) (Springer, 2003).
- Allen, W. E. et al. Thirst-associated preoptic neurons encode an aversive motivational drive. *Science* **357**, 1149–1155 (2017).
- Oka, Y., Ye, M. & Zuker, C. S. Thirst driving and suppressing signals encoded by distinct neural populations in the brain. *Nature* **520**, 349–352 (2015).
- Simpson, J. B. & Routtenberg, A. Subfornical organ: site of drinking elicitation by angiotensin II. *Science* **181**, 1172–1175 (1973).
- Smith, P. M., Beninger, R. J. & Ferguson, A. V. Subfornical organ stimulation elicits drinking. *Brain Res. Bull.* **38**, 209–213 (1995).
- Betley, J. N. et al. Neurons for hunger and thirst transmit a negative-valence teaching signal. *Nature* **521**, 180–185 (2015).
- Nation, H. L., Nicoleau, M., Kinsman, B. J., Browning, K. N. & Stocker, S. D. DREADD-induced activation of subfornical organ neurons stimulates thirst and salt appetite. *J. Neurophysiol.* **115**, 3123–3129 (2016).
- Abbott, S. B., Machado, N. L., Geerling, J. C. & Saper, C. B. Reciprocal control of drinking behavior by median preoptic neurons in mice. *J. Neurosci.* **36**, 8228–8237 (2016).
- Matsuda, T. et al. Distinct neural mechanisms for the control of thirst and salt appetite in the subfornical organ. *Nat. Neurosci.* **20**, 230–241 (2017).
- Miselis, R. R., Shapiro, R. E. & Hand, P. J. Subfornical organ efferents to neural systems for control of body water. *Science* **205**, 1022–1025 (1979).
- Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8**, 1263–1268 (2005).
- Wickersham, I. R. et al. Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron* **53**, 639–647 (2007).
- Yang, C. F. et al. Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. *Cell* **153**, 896–909 (2013).
- Roth, B. L. DREADDs for neuroscientists. *Neuron* **89**, 683–694 (2016).
- Lerner, T. N. et al. Intact-brain analyses reveal distinct information carried by SNc dopamine subcircuits. *Cell* **162**, 635–647 (2015).
- Richards, P. et al. Identification and characterization of GLP-1 receptor-expressing cells using a new transgenic mouse model. *Diabetes* **63**, 1224–1233 (2014).
- Petreanu, L., Huber, D., Sobczyk, A. & Svoboda, K. Channelrhodopsin-2-assisted circuit mapping of long-range callosal projections. *Nat. Neurosci.* **10**, 663–668 (2007).
- McKay, N. J., Galante, D. L. & Daniels, D. Endogenous glucagon-like peptide-1 reduces drinking behavior and is differentially engaged by water and food intakes in rats. *J. Neurosci.* **34**, 16417–16423 (2014).
- Betley, J. N., Cao, Z. F., Ritola, K. D. & Sternson, S. M. Parallel, redundant circuit organization for homeostatic control of feeding behavior. *Cell* **155**, 1337–1350 (2013).
- Cunningham, J. T., Beltz, T., Johnson, R. F. & Johnson, A. K. The effects of ibotenate lesions of the median preoptic nucleus on experimentally-induced and circadian drinking behavior in rats. *Brain Res.* **580**, 325–330 (1992).
- McKinley, M. J., Mathai, M. L., Pennington, G., Rundgren, M. & Vivas, L. Effect of individual or combined ablation of the nuclear groups of the lamina terminalis on water drinking in sheep. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **276**, R673–R683 (1999).
- McKinley, M. J. et al. The median preoptic nucleus: front and centre for the regulation of body fluid, sodium, temperature, sleep and cardiovascular homeostasis. *Acta Physiol. (Oxf.)* **214**, 8–32 (2015).
- Oka, Y., Butnaru, M., von Buchholtz, L., Ryba, N. J. & Zuker, C. S. High salt recruits aversive taste pathways. *Nature* **494**, 472–475 (2013).
- Yarmolinsky, D. A., Zuker, C. S. & Ryba, N. J. Common sense about taste: from mammals to insects. *Cell* **139**, 234–244 (2009).
- Zocchi, D., Wennemuth, G. & Oka, Y. The cellular mechanism for water detection in the mammalian taste system. *Nat. Neurosci.* **20**, 927–933 (2017).
- Thrasher, T. N., Keil, L. C. & Ramsay, D. J. Drinking, oropharyngeal signals, and inhibition of vasopressin secretion in dogs. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **253**, R509–R515 (1987).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank B. Ho, A. Qin and M. Liu for technical assistance, D. J. Anderson for sharing Ai110 mice, members of the Oka laboratory, and J. R. Cho for comments. We also thank N. Shah for Casp3 viruses, N. F. Dalleska, and the Beckman Institute at Caltech for technical assistance. This work was supported by Startup funds from the President and Provost of California Institute of Technology and the Biology and Biological Engineering Division of California Institute of Technology. Y.O. is also supported by the Searle Scholars Program, the Mallinckrodt Foundation, the Okawa Foundation, the McKnight Foundation and the Klingenstein-Simons Foundation, and National Institutes of Health U01 (U01 NS099717).

**Author Contributions** V.A. and Y.O. conceived the research program and designed experiments. V.A., with assistance from S.K.G., S.L. and Y.O., carried out the experiments and analysed data. B.W. and C.L. performed all slice patch-clamp recordings. T.J.D. and K.D. provided technical advice on setting up fibre photometry. F.R. and F.G. generated and provided *Glp1r-cre* mice. V.A. and S.K.G. together with Y.O. wrote the paper. Y.O. supervised the entire work.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Y.O. ([yoka@caltech.edu](mailto:yoka@caltech.edu)).

**Reviewer Information** Nature thanks M. McKinley and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**Animals.** All animal procedures were performed in accordance with the US NIH guidance for the care and use of laboratory animals and were approved by the Institutional Animal Care and Use Committee (protocol no: 1694-14, California Institute of Technology). Mice used for data collection were both males and females, at least eight weeks of age. The following mice were purchased from the Jackson Laboratory: C57BL/6J, stock number 000664; *Slc32a1-cre* (also known as *Vgat-cre*), stock number 016962; Ai9, stock number 007909; Ai3, stock number 007903; *Slc17a6-cre* (also known as *Vglut2-cre*), stock number 016963 and *Nos1-cre*, stock number 017526. *Glp1r-cre* and Ai110 lines were provided by F. Gribble (Cambridge) and D. Anderson (Caltech), respectively. Mice were housed in temperature- and humidity-controlled rooms with a 13 h:11 h light:dark cycle with *ad libitum* access to chow and water.

**Viral constructs.** The following AAVs were purchased from the UNC Vector Core: AAV1-CA-FLEX-RG,  $4 \times 10^{12}$  copies per ml; AAV1-EF1a-FLEX-TVA-mCherry,  $6 \times 10^{12}$  copies per ml; AAV2-EF1a-DIO-hChR2-eYFP,  $5.6 \times 10^{12}$  copies per ml; AAV2-hSyn-DIO-hM4D(Gi)-mCherry,  $3.7 \times 10^{12}$  copies per ml; AAV2-EF1a-DIO-mCherry,  $5.7 \times 10^{12}$  copies per ml; AAV5-CamKIIa-hM4D(Gi)-mCherry,  $4.3 \times 10^{12}$  copies per ml; AAV5-FLEX-taCasp3-TEVP,  $5.3 \times 10^{12}$  copies per ml. The following AAVs were purchased from the UPenn Vector Core: AAV1-Syn-FLEX-GCaMP6s-WPRE-SV40,  $2.9 \times 10^{13}$  genome copies per ml; AAV1-Syn-GCaMP6s-WPRE-SV40,  $2.28 \times 10^{13}$  genome copies per ml; AAV1-CamKII-eYFP-WPRE-hGH,  $1.86 \times 10^{13}$  genome copies per ml; AAV2-EF1a-DIO-eYFP-WPRE-hGH,  $3.05 \times 10^{12}$  genome copies per ml. EnvA G-deleted Rabies-eGFP ( $1.6 \times 10^8$  transduction units per ml) was purchased from the Salk Institute. Herpes simplex virus (hEF1a-LS1L-mCherry HT) was purchased from the Vector Core Facility at the Massachusetts Institute of Technology.

**Surgery.** All procedures were adopted from a previous report<sup>19</sup>. Mice were anaesthetized with a mixture of ketamine ( $1 \text{ mg ml}^{-1}$ ) and xylazine ( $10 \text{ mg ml}^{-1}$ ) in isotonic saline, injected intraperitoneally (i.p.) at  $10 \mu\text{l g}^{-1}$  bodyweight. The mice were then placed in a stereotaxic apparatus (Narishige Apparatus) on a heating pad. An incision was made to expose the skull. The three-dimensional magnetic resonance imaging coordinate system was used to align the skull reference. A small craniotomy, less than 1 mm, was made using a hand drill at the regions of interest. Viral constructs were injected using a pressure injection system (Nanoliter 2000) using a pulled glass capillary at  $100 \text{ nl min}^{-1}$ . The coordinates were: anteroposterior  $-4,030$ , mediolateral  $0$ , dorsoventral  $-2,550$  (200-nl injection) for the SFO; anteroposterior  $-3,100$ , mediolateral  $0$ , dorsoventral  $-4,080$  (100-nl injection) and  $-3,800$  (50–100-nl injection) for the MnPO; and anteroposterior  $-2,700$ , mediolateral  $0$ , dorsoventral  $-4,900$  (75-nl injection) for the OVLT. For optogenetic implants, a 200- $\mu\text{m}$  fibre bundle (FT200EMT, Thorlabs) glued to a ceramic ferrule (Thorlabs) with epoxy was used. For photometry implants, a 400- $\mu\text{m}$  fibre bundle (BFH48-400, Thorlabs) glued to a ceramic ferrule with low autofluorescence epoxy (EPO-TEK301) or a custom-made implant (Doric Lenses) was used. A fibre was implanted 200–300  $\mu\text{m}$  (for photostimulation) or 0–50  $\mu\text{m}$  (for photometry) above the virus injection site. After the application of a local anaesthetic to the sides of the skin incision, the implants were permanently fixed to the skull using dental cement. Cannulated mice were placed in a clean cage on a heating pad to recover from anaesthesia. Mice were kept in their home cage for at least ten days before any behavioural tests.

**Photostimulation.** For optogenetic experiments, photostimulation was performed using 473-nm laser pulses: 20 ms, 5 Hz (for OVLT) or 20 Hz (for SFO and MnPO) delivered via a custom-made optic cable using a pulse generator (World Precision Instruments). The laser intensity was maintained at 5 mW (for OVLT) or 10 mW (for SFO and MnPO) at the tip of the fibre.

**Behavioural assays.** For water-restriction experiments, mice were provided with 1 ml of water daily. For food-restriction experiments, mice were provided with 0.5 pellets per 20 grams of body weight daily. All assays were performed in a modified lickometer as described previously<sup>39</sup> or a Biodaq monitoring system (Research Diets Inc.). For all photometry assays, mice were acclimatized for 10–15 min in the lickometer cage before stimuli were given.

**Long-term access assays.** For optogenetic testing (Fig. 1g and Extended Data Fig. 2c, e), satiated mice were given *ad libitum* access to water with photostimulation. Photostimulation was delivered for 1 s at 3-s intervals throughout the behavioural sessions. For Fig. 2e and Extended Data Fig. 6c, mice were given access to water for 20 min after 24-h water restriction, and photostimulation was delivered for the first 10 min. For feeding assays (Fig. 2e), mice were single-housed in Biodaq cages after 24-h food restriction, and chow intake was measured for 20 min with or without light stimulation. For acute inhibition experiments, mice were given access to 150 mM NaCl (Fig. 5b and Extended Data Fig. 8c) or water (Fig. 1g and Extended Data Figs 2e, 8a, b, d) for 20–30 min after 24 h water restriction, or 300 mM sucrose (Fig. 1g and Extended Data Fig. 2e) after food restriction. For all

acute inhibition experiments, CNO was injected at  $10 \text{ mg kg}^{-1}$  body weight, 30 min before the start of the behaviour session. For acute activation experiments, CNO was injected at  $1 \text{ mg kg}^{-1}$  body weight (Extended Data Fig. 7e), 30 min before the start of the behaviour session. For Fig. 3a and Extended Data Figs 4a and 8f, access to water or saline was provided for 30 min after 24 h of water restriction. For Fig. 4a, water or HydroGel (ClearH<sub>2</sub>O) in a cup was provided for 30 min after 24 and 36 h of water restriction, respectively. The weight of the cup was measured before and after the behaviour session. For Fig. 4d, 0.5 pellets of chow was provided for 30 min after 24 h of food restriction. The entire session was recorded using a camera at 30 frames per second, and ingestion episodes were manually annotated.

**Salt- or mannitol-loading experiments.** 150  $\mu\text{l}$  or 300  $\mu\text{l}$  of 2 M NaCl, or 300  $\mu\text{l}$  of 2 M mannitol, was injected intraperitoneally at the end of the acclimatization period. For Fig. 1i and Extended Data Fig. 2f, CNO or vehicle (water) was injected 10 min before the injection of NaCl or mannitol.

**Brief access assays.** For optogenetic experiments, behavioural assays were performed essentially as previously described<sup>19</sup>. Satiated mice were tested in a gustometer for 10–15 trials (Fig. 1e and Extended Data Fig. 1d). The laser pulses were delivered for 20 s of the 40-s trial. After the first lick, mice were given access to a water spout for 5 s. For photometry recording (Fig. 3b and Extended Data Fig. 7a), water-restricted mice were presented with one of the following four stimuli for 30 s: water, isotonic saline, silicone oil or empty bottle (control). Under food-restricted conditions (Fig. 3c and Extended Data Fig. 7b), a bottle containing 300 mM sucrose, peanut butter coated on a spout, or an empty bottle was presented for 30 s. To avoid the effect of internal state changes, we used the data from the first stimulus presentation in each session. To test the effect of temperature (Fig. 4f), three bottles of water at 4 °C, room temperature (25 °C) or 37 °C were placed at the start of the acclimatization period (10 min). Each trial was 30 s long with an inter-trial interval of 2 min. For Fig. 4e, water-restricted mice had access to water for 2 s repeated 15 times or for one 30-s period. Each presentation was followed by a 30-s interval.

**Fibre photometry.** We measured bulk fluorescence signals using fibre photometry as previously described<sup>31</sup>. In brief, 490 nm and 405 nm light-emitting diodes (Thorlabs, M490F1 and M405F1) were collimated and delivered to the brain. The light intensity was maintained at less than 100  $\mu\text{W}$  during all recordings. The fluorescence signal was then focused onto a femtowatt photoreceiver (Newport, Model 2151). The modulation and demodulation were performed with an RP2.1 real time processor (Tucker-Davis Technologies) running custom software. The licks from the lickometer were simultaneously recorded as real-time transistor–transistor logic signals to the RP2.1. Fluorescence changes were analysed using custom MATLAB (MathWorks) code as described previously<sup>31</sup>. Data were extracted and subjected to a low-pass filter at 1.8 Hz. A linear function was used to scale up the 405-nm channel signal to the 490-nm channel signal to obtain the fitted 405-nm signal. The resultant  $\Delta F/F$  was calculated as (raw 490 nm signal – fitted 405 nm signal)/(fitted 405 nm signal). For brief access tests, the area under the curve ( $\Sigma \Delta F_{\text{during}}$ ) was quantified by integrating the fluorescence signals during the bout. For all bouts, the mean fluorescence for 30 s before the first lick was calculated and subtracted from the entire session.  $\Delta F$  changes ( $\Delta F_{\text{post}} - \Delta F_{\text{pre}}$ ) were calculated by subtracting the mean fluorescence signal during the 2-s period before the first lick from the mean signal during the 2-s period at 1 min after the bout. To display traces, the fluorescence data was time-binned by a factor of  $2.5 \times$  the sampling frequency and down-sampled to 1 Hz. For long-term tests, the area under the curve was calculated for 2.5 min after the start of the bout. Changes in  $\Delta F$  were calculated by subtracting the mean signal during the 2-s period before the first lick or NaCl injection from the mean signal during the 2-s period at 5 or 10 min after the bout (Extended Data Fig. 4). For peristimulus time histograms (Fig. 4c, d), the first bout at the start of the session and the last bout within 10 min of access were used. The areas under the curve for the peristimulus time histograms were calculated during the first or the last 15 s.

**Viral tracing.** *Monosynaptic rabies tracing.* 150 nl of a mixture of AAV1-CA-FLEX-RG and AAV1-EF1a-FLEX-TVA-mCherry (4:1 ratio) was injected to the target area. Two weeks later, 200 nl of EnvA G-deleted Rabies-eGFP was injected into the same area. The mice were euthanized a week later and their brains collected.

**HSV tracing.** 200 nl of a mixture of AAV1-Syn-GCaMP6s-WPRE-SV40 and hEF1-LS1L-mCherry HT (2:5 ratio) was injected to the SFO of *Vgat-cre* mice. The GCaMP virus was used to mark the injection site. The mice were euthanized three weeks later and their brains collected.

The sections were imaged using a confocal microscope (TCS SP8, Leica) or a slide scanner (VS120, BX61VS, Olympus) at  $20 \times$ . The slide scanner images were used to count cells using ImageJ. Representative images in Figs 1a, 2a and Extended Data Fig. 5 are from the confocal microscope. Regions with an average greater than 10 rabies-virus-positive cells were included in the analysis.

**Histology.** Mice were deeply anaesthetized with carbon dioxide and then transcardially perfused with PBS followed by 4% paraformaldehyde in PBS (pH 7.4) at 4 °C.



The brains were extracted and fixed in 4% paraformaldehyde at 4 °C overnight. 100 µm coronal sections were prepared using a vibratome (Leica, VT-1000 s) for antibody staining. The primary antibodies (1:500 dilution) used were: goat anti-c-Fos (Santa Cruz, SC-52G), rabbit anti-NOS1 (Santa Cruz, sc-648), rabbit anti-GAD65<sup>+</sup>GAD67 (Abcam, ab183999), chicken anti-GFP (Abcam, ab13970) and rat anti-mCherry (Thermo Fisher, M11217). After washing three times with PBS, the sections were incubated with secondary antibodies (1:500 dilution) in blocking buffer for 4 h. The GAD65/67 primary/secondary antibody incubation solution was prepared without detergent. Fluorescence *in situ* hybridization was carried out using the RNAscope fluorescent multiplex kit (Advanced Cell Diagnostics) in accordance with the manufacturer's instructions. *Glp1r-cre/Ai9* mice were used with probes targeted to tdTomato and GLP1R.

**RNA sequencing analysis.** The dorsal lamina terminalis in *Vgat-cre/Ai9* mice were dissected under a fluorescence microscope. To minimize contamination from other tissues, the lamina terminalis tissue containing the SFO and dorsal MnPO were peeled off. For non-lamina-terminalis control, we dissected small tissues of the cortex from the same mice. These samples were dissociated into single cells using the Papain Dissociation System (Worthington), labelled with 4',6-diamidino-2-phenylindole (DAPI) and the tdTomato-positive neurons were sorted using a flow cytometer (MoFlo Astrios, Beckman Coulter). RNA was extracted using a PicoPure RNA isolation kit (Applied Biosystems) and complementary DNA was prepared using an Ovation RNA-seq V2 kit (Nugen). Relative gene expression (Fig. 2b) was calculated as a ratio of fragments per kilobase million of the dorsal lamina terminalis to that of the cortex. The genes with fragments per kilobase million < 0.1 in the cortex were omitted from the plot.

**Slice electrophysiology.** Procedures for the preparation of acute brain slices and recordings with optogenetic stimulations were similar to those described previously<sup>19,43</sup>. After decapitation, the brain was removed and immersed in ice-cold solution. Coronal slices (300 µm) were cut using a vibratome (VT-1200 s, Leica) and moved into HEPES holding solution (in mM: NaCl 92, KCl 2.5, NaH<sub>2</sub>PO<sub>4</sub> 1.2, NaHCO<sub>3</sub> 30, HEPES 20, glucose 25, Na-ascorbate 5, thiourea 2, Na-pyruvate 3, MgSO<sub>4</sub> 2, CaCl<sub>2</sub> 2, at pH 7.35). The slices were allowed to recover at 33 °C for 30 min and then held at room temperature (around 25 °C) until use.

While recording, slices were perfused continuously (around 2 ml min<sup>-1</sup>) with artificial cerebrospinal fluid (in mM: NaCl 124, KCl 2.5, NaH<sub>2</sub>PO<sub>4</sub> 1.2, NaHCO<sub>3</sub> 24, glucose 25, MgSO<sub>4</sub> 1, CaCl<sub>2</sub> 2) at 25 °C. Neurons were visualized and targeted using an upright infrared differential interference contrast microscope (BX51WI, Olympus). Whole-cell recordings were achieved using glass pipettes with an impedance of 4–6 MΩ when filled with intracellular solution (for voltage clamp, in mM: CsCl 145, NaCl 2, HEPES 10, EGTA 0.2, QX-314 Chloride 5, Mg-ATP 4, Na-GTP 0.3, at pH 7.25; for current clamp, in mM: K-gluconate 145, NaCl 2, KCl 4, HEPES 10, EGTA 0.2, Mg-ATP 4, Na-GTP 0.3, at pH 7.25). Electrical signals were sampled at 20 kHz and filtered at 2.9 kHz using an EPC 10 system (HEKA Elektronik). To evaluate postsynaptic currents evoked by light pulses, the membrane potential of SFO<sup>nNOS</sup> (transduced with CamKII-mCherry/eYFP) or SFO<sup>non-nNOS</sup> neurons was held at -60 mV. Light pulses were generated by a mercury lamp, filtered by an optical filter (Chroma) and controlled by an electronic shutter driver (VCM-D1, UNIBLITZ). 2-ms light pulses were delivered at 1 Hz four

times, followed by a 4-s interval. We repeated this stimulus cycle 20 times. To confirm that the postsynaptic currents recorded were GABAergic, picrotoxin (150 µM) was applied through the bath for part of the experiments. To confirm glutamatergic postsynaptic currents, 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX, 10 µM) and 2-amino-5-phosphonopivalic acid (DL-APV, 25 µM) were applied through the bath. Monosynaptic connection was defined by synaptic inhibitory or excitatory postsynaptic currents with latencies less than 16.4 ms. For hM4Di experiments, current-clamp recordings were performed by applying a constant supra-threshold current injection to produce tonic action potentials. CNO (around 6 µM) was applied using a puff (30 s) from another glass pipette placed approximately 50 µm from the recorded cell.

**Plasma Na<sup>+</sup> and osmolality measurements.** After the injection of 150 µl of 2 M NaCl or 300 µl of 2 M mannitol, trunk blood was collected from wild-type mice. Plasma was then extracted after centrifugation at 1500g for 20 min. Plasma osmolality was measured using a vapour pressure osmometer (Vapro 5520). Plasma Na<sup>+</sup> concentration was measured using Dionex (Thermo) ICS 2000.

**Intra-cranial drug delivery.** 100 ng of exendin-4 (Sigma Aldrich) dissolved in 1 µl of artificial cerebrospinal fluid was delivered using a custom-made cannula and tubing (PlasticsOne) connected to a Hamilton syringe driven by a pump (NewEra PumpSystems) at 100 nl min<sup>-1</sup> into the MnPO of water-deprived mice under head-fixed conditions. Two minutes after infusion, freely moving mice were given access to water for the next 45 min. The cannula position was verified by infusing exendin-4-FAM (Anaspec) conjugate before euthanasia.

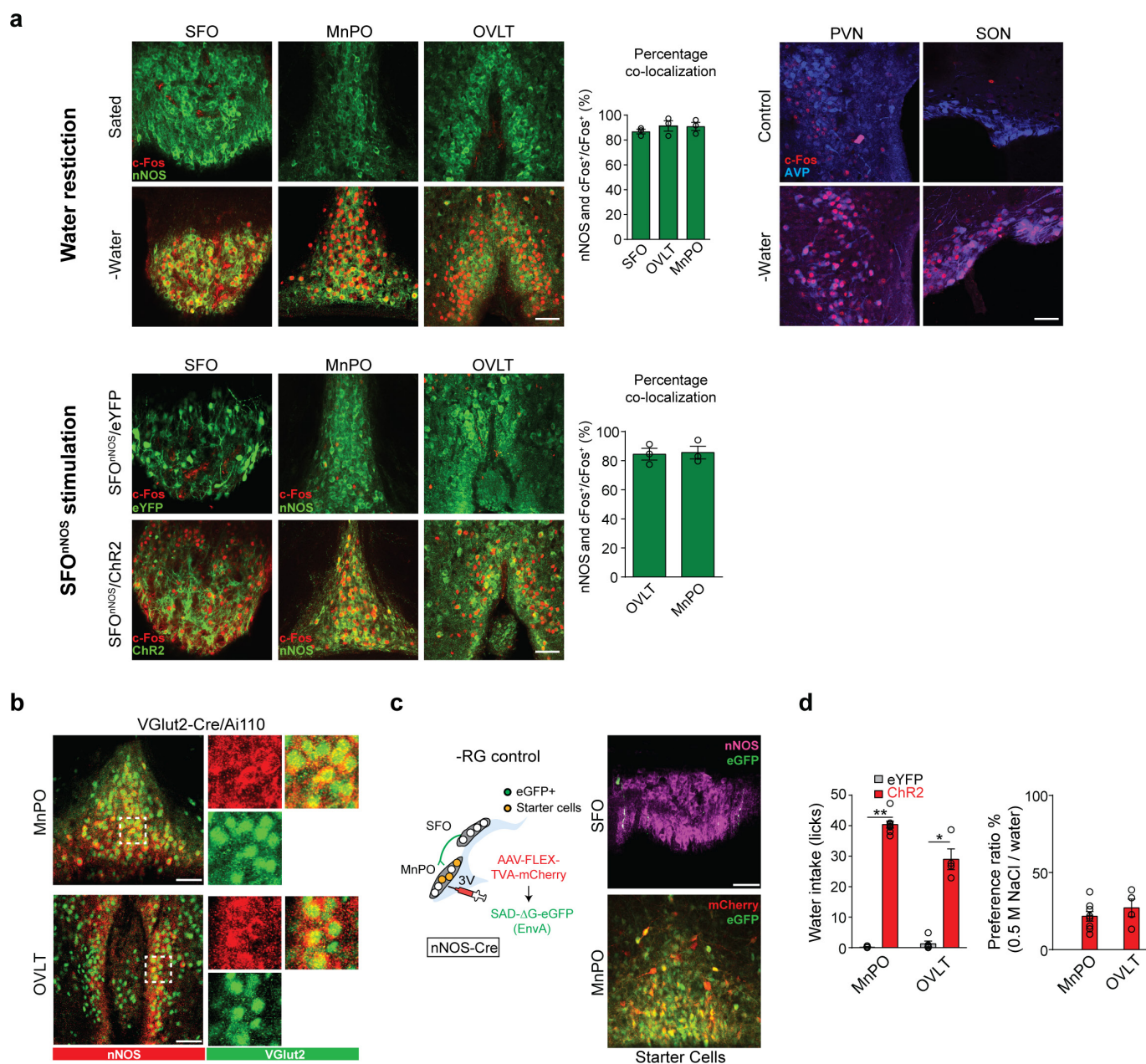
**Enzyme-linked immunosorbent assay.** Total plasma GLP1 was measured using EZGLP1T-36k kit (Millipore) as described previously<sup>44</sup>. In brief, after blood was collected in EDTA-coated tubes, plasma was isolated by centrifugation at 1500g for 20 min. Samples were then kept at -80 °C until measurement. For food-repleted (FD + F) and water-repleted (WD + W) conditions, mice were given access to Ensure for 30 min or water for 5 min, respectively.

**Statistics.** All statistical analyses were carried out using Prism (GraphPad). We used a two-tailed Mann–Whitney *U* test, a paired *t*-test or a Kruskal–Wallis one-way ANOVA, depending on the experimental paradigm. \**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001. Details of the tests used are outlined in Supplementary Table 1. No statistics to determine sample size, blinding or randomization methods were used. Viral expression and implant placement was verified by histology before mice were included in the analysis. These criteria were pre-established.

**Code availability.** Custom MATLAB code used in this study is available from the corresponding author upon reasonable request.

**Data availability.** Data are available from the corresponding author upon reasonable request.

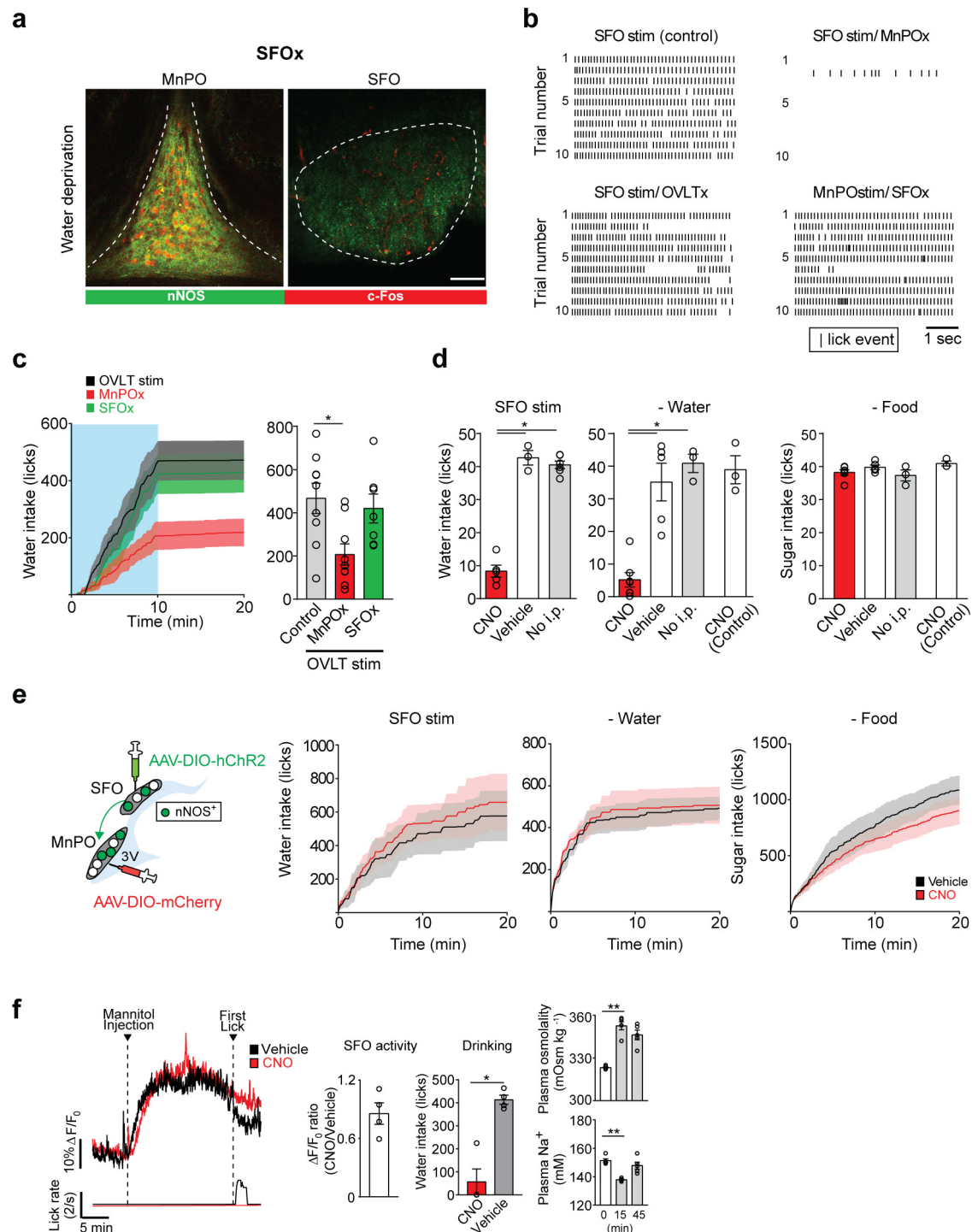
43. Krashes, M. J. *et al.* An excitatory paraventricular nucleus to AgRP neuron circuit that drives hunger. *Nature* **507**, 238–242 (2014).
44. Kahles, F. *et al.* GLP-1 secretion is increased by inflammatory stimuli in an IL-6-dependent manner, leading to hyperinsulinemia and blood glucose lowering. *Diabetes* **63**, 3221–3229 (2014).
45. Paxinos, G. & Franklin, K. B. J. *The Mouse Brain in Stereotaxic Coordinates* 2nd edn (Academic, 2001).



### Extended Data Figure 1 | Optogenetic activation MnPO<sup>nNOS</sup> and OVLT<sup>nNOS</sup> neurons induces robust water intake in satiated mice.

**a**, Water restriction (top) and SFO<sup>nNOS</sup> photostimulation (bottom) induces robust c-Fos expression in the SFO, MnPO and OVLT, compared to control conditions. A majority of c-Fos signals in these areas overlapped with nNOS-expressing neurons. The graph shows the quantification of the overlap between nNOS and c-Fos signals ( $n = 3$  mice). c-Fos signals in the paraventricular nucleus (PVN) and supraoptic nucleus (SON) overlapped with vasopressin (AVP)-expressing neurons. **b**, MnPO (top) and OVLT (bottom) excitatory neurons visualized in VGlut2/Ai110 transgenic mice co-stained with nNOS (red, antibody staining). MnPO<sup>nNOS</sup> and OVLT<sup>nNOS</sup> neurons co-express a glutamatergic marker.  $92.2 \pm 4.9\%$  of nNOS-expressing neurons were excitatory, and  $80.9 \pm 2.6\%$  of excitatory neurons

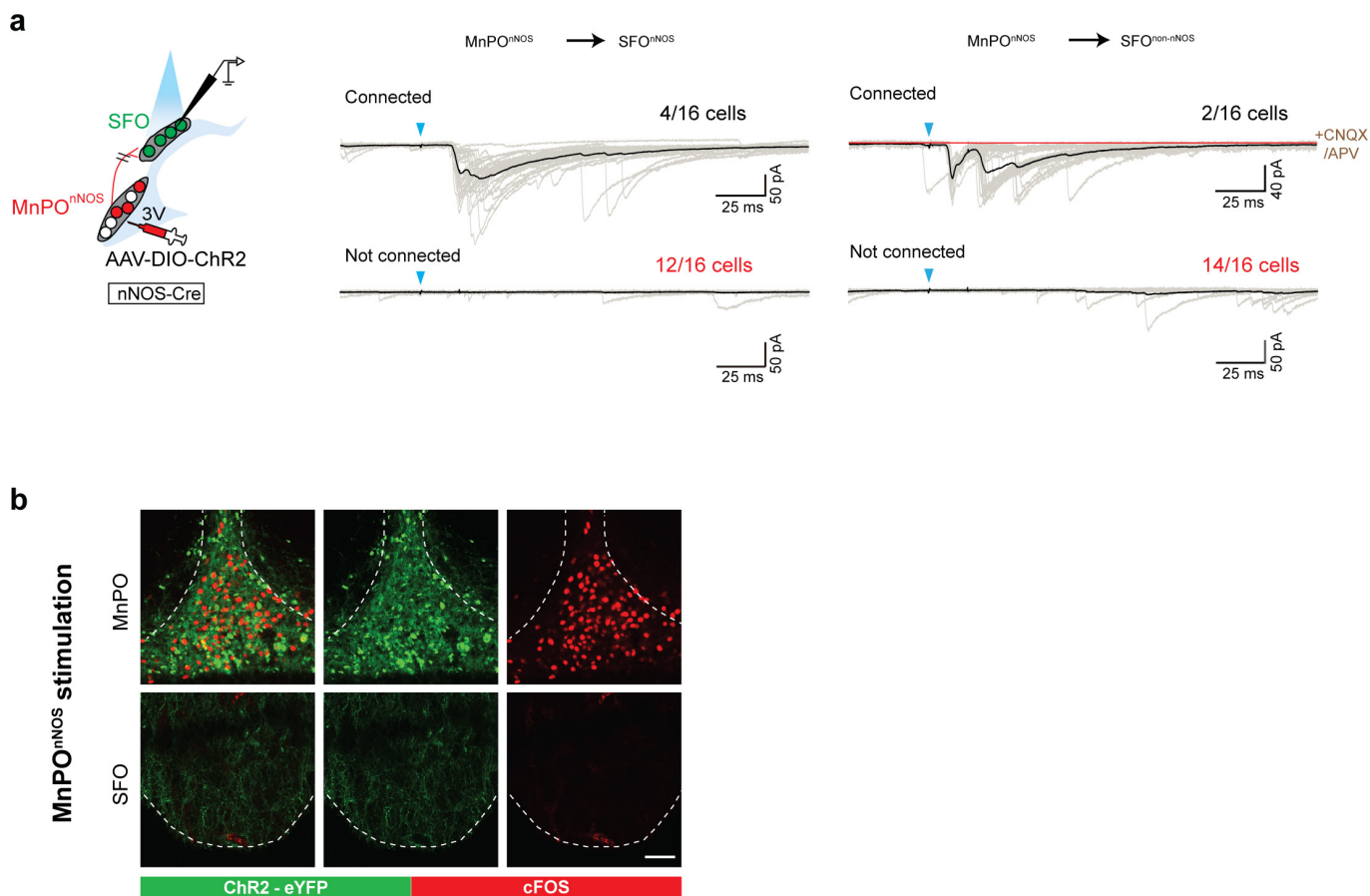
are nNOS-expressing in the MnPO ( $n = 3$  mice). Magnified images are shown on the right. **c**, Left, scheme of the control experiments for monosynaptic rabies tracing. Right, a representative image of the MnPO of an nNOS-cre mouse transduced with AAV-EF1a-FLEX-TVA-mCherry (red) followed by EnvA G-deleted Rabies-eGFP (bottom). No eGFP<sup>+</sup> cells were present in the SFO (top, one of two mice). **d**, Photostimulation of ChR2-expressing MnPO<sup>nNOS</sup> and OVLT<sup>nNOS</sup> neurons (red bars,  $n = 8$  and 4 mice for MnPO and OVLT respectively) triggered intense drinking; control mice infected with AAV-DIO-eYFP showed no such response (grey bars,  $n = 5$  mice). Photostimulated mice showed a strong preference for water over a highly concentrated NaCl solution (500 mM, right panel).  $*P < 0.05$ ,  $**P < 0.01$ ; by two-tailed Mann-Whitney  $U$  test. All error bars show mean  $\pm$  s.e.m. Scale bars, 50  $\mu$ m.



**Extended Data Figure 2 | MnPO<sup>nNOS</sup> neurons are necessary for the induction of drinking by SFO<sup>nNOS</sup> photostimulation.** **a**, Casp3-TEVp efficiently eliminates SFO<sup>nNOS</sup> neurons (right) without affecting MnPO<sup>nNOS</sup> neurons (left). **c**-Fos expression pattern is shown after water-restriction (red). **b**, Raster plots representing licking events during the 5-s session with photostimulation. **c**, Ablation of MnPO<sup>nNOS</sup> (MnPOx) but not SFO<sup>nNOS</sup> (SFOx) neurons attenuated the drinking response to OVLT<sup>nNOS</sup> photostimulation (left, 10 min, blue box). Quantification of the number of licks during the 10-min light-on period (right,  $n = 9$  mice for controls and MnPOx and  $n = 7$  mice for SFOx). **d**, 5-s brief-access assays to examine the necessity of MnPO<sup>nNOS</sup> neurons. Acute inhibition of MnPO<sup>nNOS</sup> neurons by CNO injection severely reduced SFO<sup>nNOS</sup>-stimulated (left,  $n = 5$  mice for CNO,  $n = 3$  mice for vehicle, and  $n = 6$  mice for no i.p.) and dehydration-induced water intake (middle,  $n = 7$  mice for CNO,  $n = 5$  mice for vehicle, and  $n = 3$  mice for no i.p.). However, the same treatment did not suppress sucrose consumption (300 mM, right,

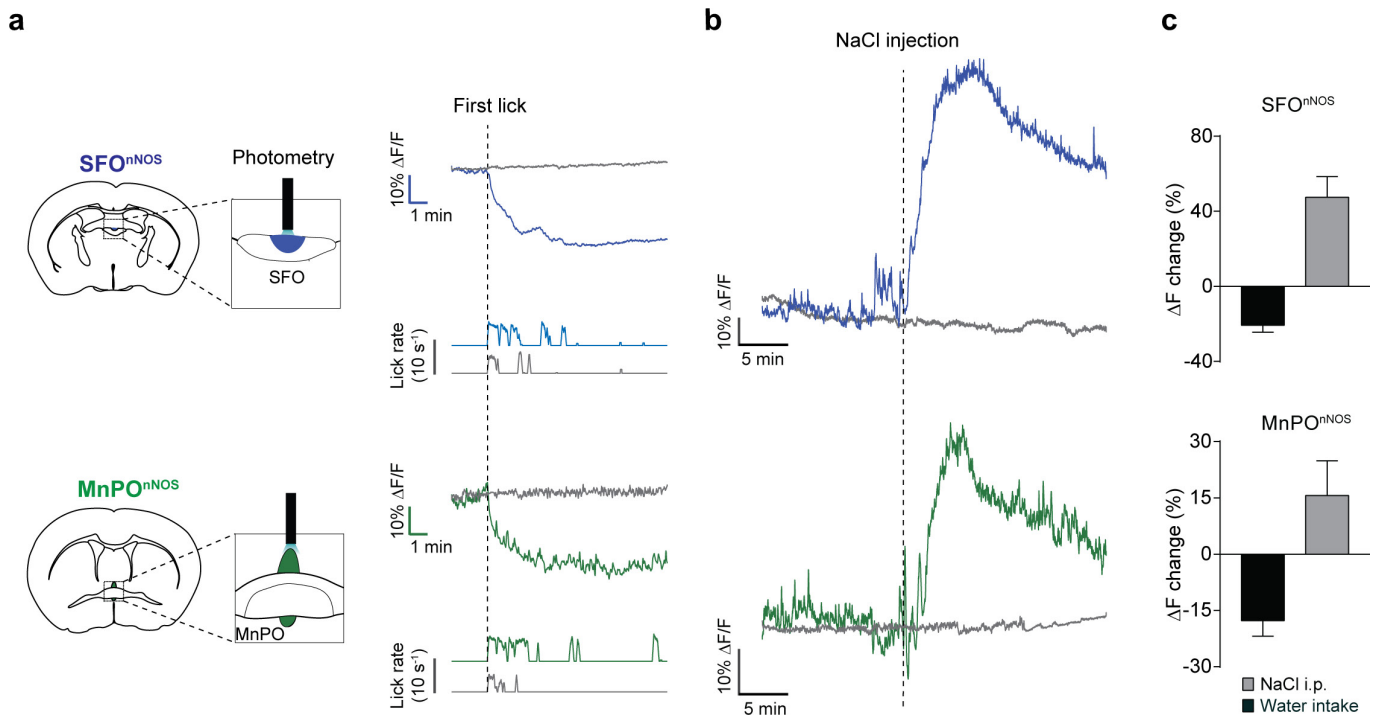
$n = 6$  mice for CNO,  $n = 5$  mice for vehicle, and  $n = 3$  mice for no i.p.). Control mice transduced by AAV-DIO-mCherry in the MnPO showed no reduction after water or food-restriction ( $n = 3$  mice). **e**, mCherry control for Fig. 1g. Cumulative water intake in *nNOS*-cre mice transduced with AAV-DIO-mCherry in the MnPO, AAV-DIO-ChR2-eYFP in the SFO under photostimulated (left,  $n = 5$  mice) or water-restricted conditions (middle,  $n = 6$  mice), and sucrose (300 mM) intake under food-restricted conditions (right,  $n = 5$  mice). **f**, Intraperitoneal injection of mannitol robustly activated SFO<sup>nNOS</sup> neurons with (red trace) or without (black trace) CNO injection (left). CNO injection drastically suppressed drinking behaviour without changing the activity of SFO<sup>nNOS</sup> neurons (middle,  $n = 4$  mice). Plasma osmolality was increased by the injection of mannitol (right,  $n = 5$  mice). \* $P < 0.05$ , \*\* $P < 0.01$ , by paired two-tailed *t*-test or Kruskal–Wallis one-way ANOVA test with Dunn's correction for multiple comparisons. All error bars and shaded areas show mean  $\pm$  s.e.m. Scale bar, 50  $\mu$ m.





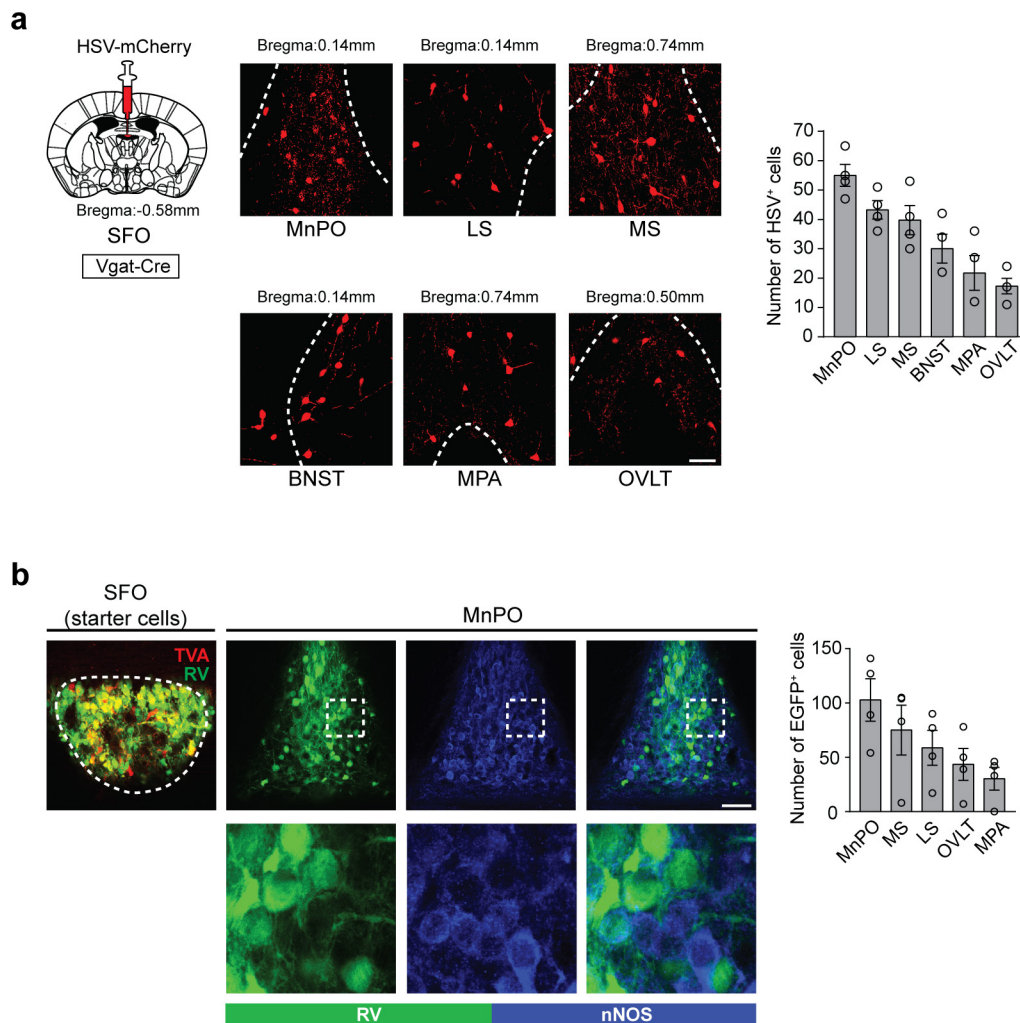
**Extended Data Figure 3 | The SFO receives sparse monosynaptic input from MnPO<sup>nNOS</sup> neurons.** **a**, Left, schematic for the assessment of the MnPO<sup>nNOS</sup> → SFO monosynaptic connection (left). Right, whole-cell patch-clamp recording from SFO neurons was performed with optogenetic stimulation of MnPO<sup>nNOS</sup> → SFO projections. Excitatory synaptic currents were measured in the presence (red trace) or absence (black trace) of CNQX (10 μM) + DL-APV (25 μM) after photostimulation (2 ms, blue

arrowheads). Most SFO<sup>nNOS</sup> neurons (12 out of 16 cells, labelled with mCherry, middle panel) or SFO<sup>non-nNOS</sup> neurons (14 out of 16 cells, right panel) did not receive monosynaptic input from MnPO<sup>nNOS</sup> neurons. **b**, Representative image (one out of three mice) of robust c-Fos expression (red) in the MnPO (top) but not in the SFO (bottom) by photostimulation of ChR2 expressing MnPO<sup>nNOS</sup> neurons. Scale bar, 50 μm.



**Extended Data Figure 4 | Neural dynamics of SFO<sup>nNOS</sup> and MnPO<sup>nNOS</sup> neurons.** **a**, Left, Schematic of fibre photometry experiments from SFO<sup>nNOS</sup> (top) and MnPO<sup>nNOS</sup> (bottom) neurons. *nNOS-cre* mice were injected with AAV-FLEX-GCaMP6s or eYFP into the SFO and MnPO. Right, representative traces showing the real-time activity of the SFO<sup>nNOS</sup> (blue trace) and MnPO<sup>nNOS</sup> (green trace) populations with water intake in water-restricted mice. Grey traces show the activity of eYFP control mice. Corresponding lick patterns are also shown (lower traces). SFO<sup>nNOS</sup>

and MnPO<sup>nNOS</sup> neurons are rapidly and persistently inhibited by water drinking. **b**, SFO<sup>nNOS</sup> and MnPO<sup>nNOS</sup> neurons are sensitive to thirst-inducing stimuli. Intraperitoneal injection of NaCl (2 M, 300 μl) in a water-satiated animal robustly activated SFO<sup>nNOS</sup> (blue) and MnPO<sup>nNOS</sup> (green) neurons. **c**, Quantification of the neuronal responses. During liquid intake (black bars,  $n = 4$  mice for SFO,  $n = 6$  mice for MnPO) and sodium loading (grey bars,  $n = 5$  mice), both SFO<sup>nNOS</sup> and MnPO<sup>nNOS</sup> neurons showed opposite activity changes. All error bars show mean  $\pm$  s.e.m.

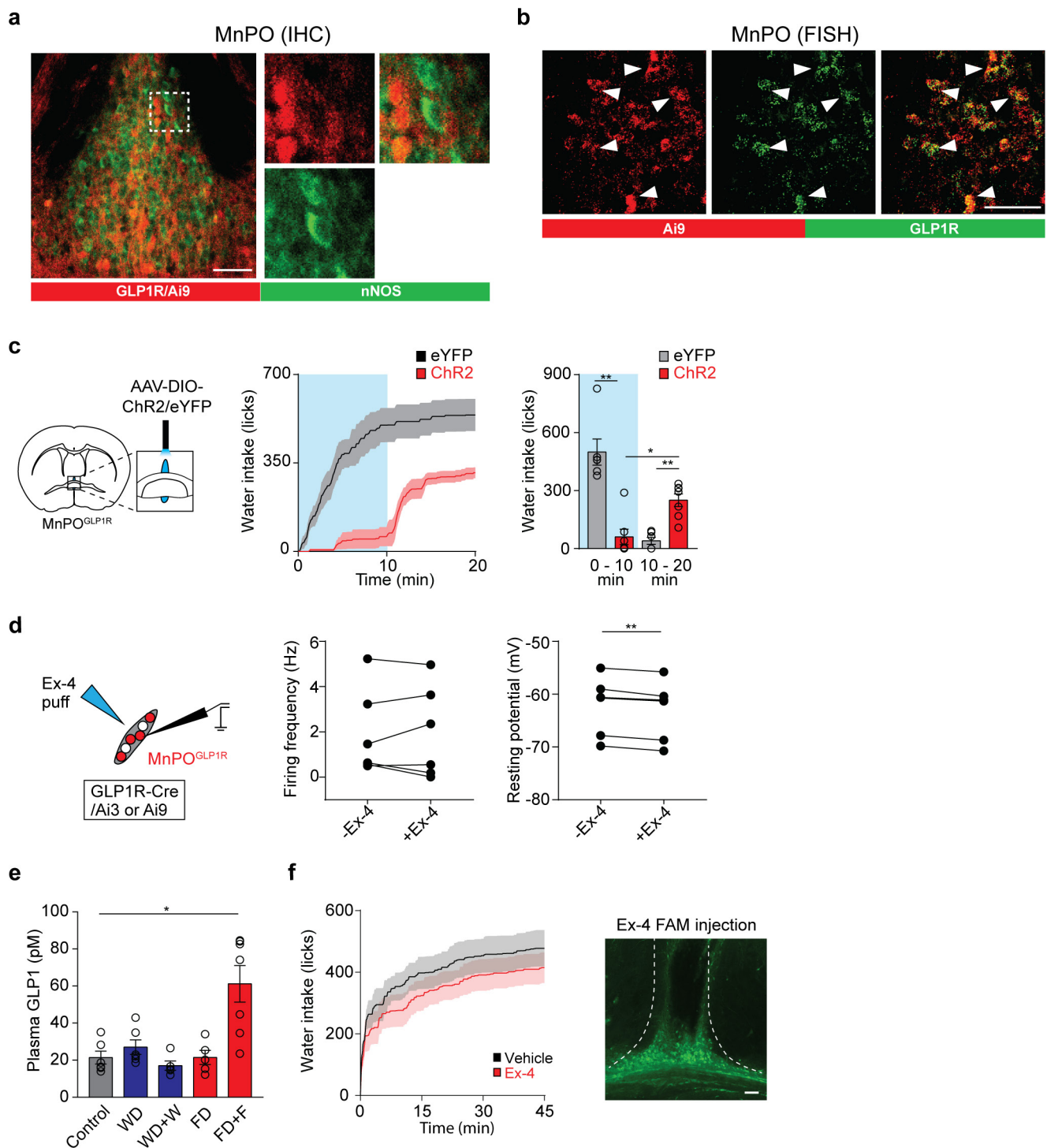


### Extended Data Figure 5 | Mapping of inhibitory inputs to the SFO.

**a**, Left, a schematic for retrograde tracing of inhibitory inputs to the SFO by HSV-mCherry. Shown are the major inhibitory inputs to the SFO. Right, quantification of HSV-positive neurons ( $n = 4$  mice). LS, lateral septum; MS, medial septum; BNST, bed nucleus of the stria terminalis; MPA, medial preoptic area. **b**, Monosynaptic retrograde rabies tracing of SFO<sup>nNOS</sup> neurons. Left, a representative image of the SFO of an *nNOS-cre* mouse transduced with AAV-CA-FLEX-RG and AAV-EF1a-FLEX-

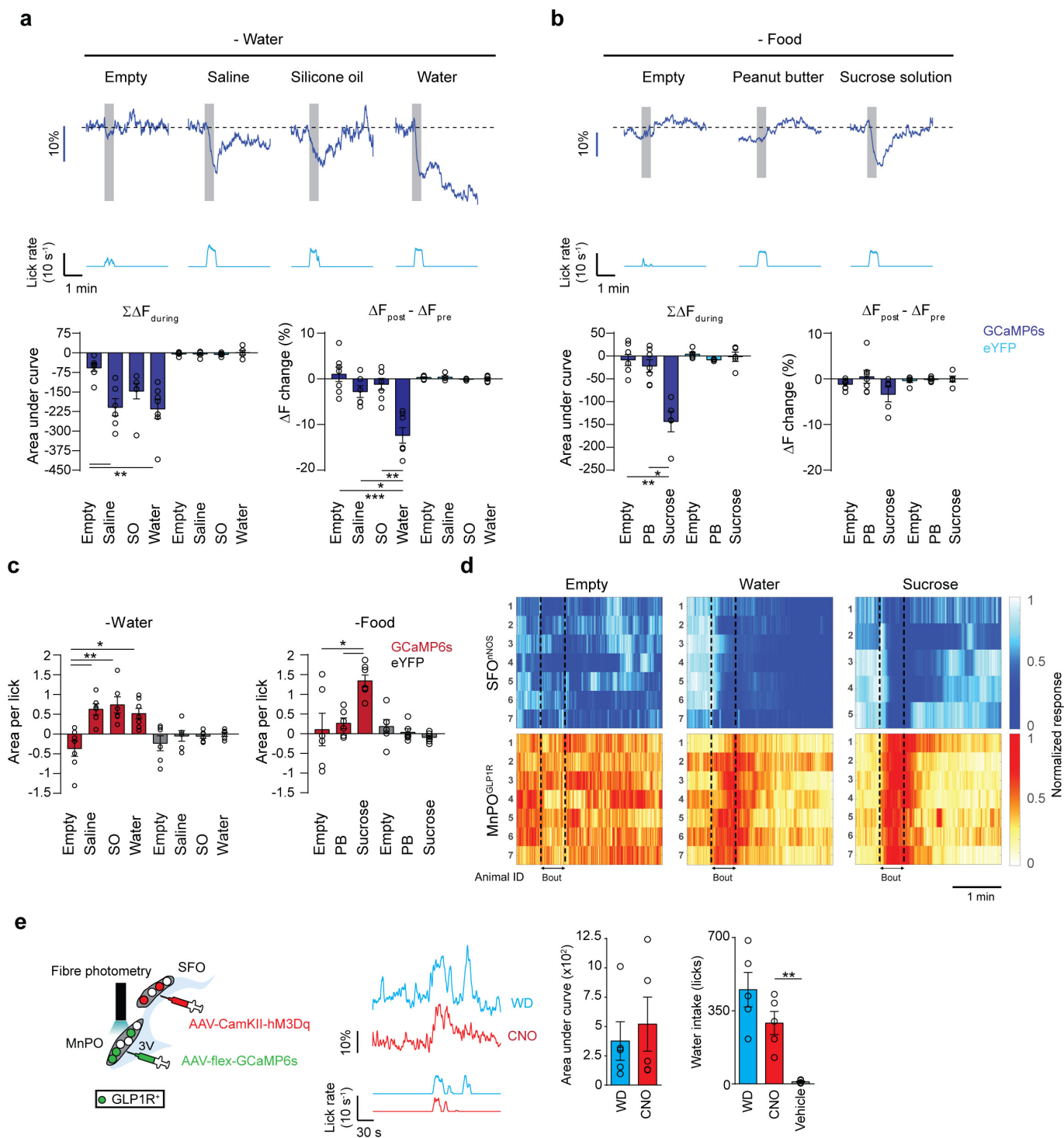
TVA-mCherry followed by EnvA G-deleted Rabies-eGFP. Right, almost no eGFP-positive neurons in the MnPO (green,  $5.4 \pm 1.3\%$ ,  $n = 4$  mice) overlapped with excitatory *nNOS*-expressing neurons (blue). Maximum inputs to the SFO<sup>nNOS</sup> neurons are from the MnPO, followed by the MS, LS, MPA and OVLT ( $n = 4$  mice). All error bars show mean  $\pm$  s.e.m. Scale bars,  $50 \mu\text{m}$ . The mouse brain in this figure has been reproduced from the mouse brain atlas<sup>45</sup>.





**Extended Data Figure 6 | The MnPO<sup>GLP1R</sup> population does not overlap with nNOS-expressing neurons.** **a**, nNOS antibody staining (green) of the MnPO from a *Glp1r-cre*/Ai9 transgenic mouse expressing tdTomato in MnPO<sup>GLP1R</sup> neurons (red). No substantial overlap was observed between these populations ( $4.3 \pm 0.9\%$  of GLP1R-expressing neurons,  $n = 3$  mice). **b**, Fluorescence *in situ* hybridization (FISH) shows that a majority of Ai9 expression (red,  $91.9 \pm 2.4\%$ ,  $n = 3$  mice) closely overlaps with endogenous GLP1R expression (green). **c**, Left, a diagram showing optogenetic stimulation of MnPO<sup>GLP1R</sup> neurons transduced with AAV-DIO-ChR2-eYFP or AAV-DIO-eYFP. Right, stimulation of ChR2-expressing MnPO<sup>GLP1R</sup> neurons inhibited drinking after water restriction as compared to eYFP controls ( $n = 7$  mice for ChR2,  $n = 6$  mice for controls, blue box indicates the Light-ON period). For statistical analysis, we used the same dataset as for 0–10 min from Fig. 2e. **d**, GLP1 has minor effects on acute drinking behaviour. A diagram of whole-cell recording

from MnPO<sup>GLP1R</sup> neurons is shown on the left. A GLP1 agonist, exendin-4 (Ex-4), had no effect on the firing frequency of MnPO<sup>GLP1R</sup> neurons in brain slice preparation (middle,  $n = 6$  neurons). However, there was a small decrease in the resting membrane potential (right,  $n = 6$  neurons). **e**, Enzyme-linked immunosorbent assay analysis of plasma GLP1 levels. Feeding behaviour induced robust plasma GLP1 secretion whereas water intake did not ( $n = 5$  mice for WD + W and FD,  $n = 6$  mice for control and WD, and  $n = 7$  mice for FD + F). **f**, Left, intra-cranial injection of Ex-4 (red trace,  $n = 7$  mice) into the MnPO had no effect on water intake after water deprivation as compared to vehicle injection (artificial cerebrospinal fluid, black trace,  $n = 7$  mice). Right, a representative injection pattern visualized with fluorescent Ex-4 FAM.  $*P < 0.05$ ,  $**P < 0.01$ , two-tailed Mann–Whitney *U* test or paired *t*-test or Kruskal–Wallis one-way ANOVA test with Dunn's correction for multiple comparisons. All error bars and shaded areas show mean  $\pm$  s.e.m. Scale bars,  $50 \mu\text{m}$ .

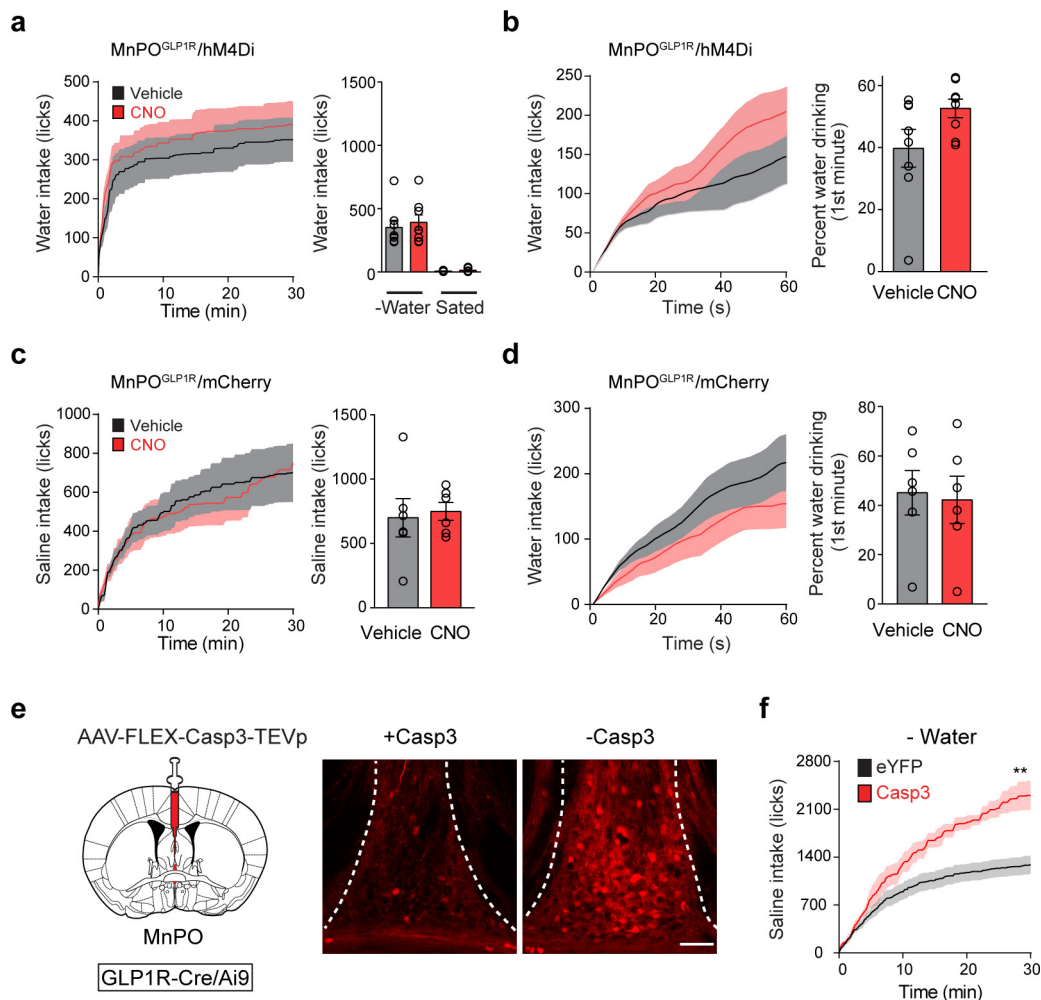


Extended Data Figure 7 | See next page for caption.

**Extended Data Figure 7 | *In vivo* activation patterns of MnPO<sup>GLP1R</sup> and SFO<sup>nNOS</sup> neurons upon ingestion.** **a**, SFO<sup>nNOS</sup> neurons are negatively and chronically regulated by water drinking. Representative responses of SFO<sup>nNOS</sup> (blue traces) to different types of liquids under water-restricted conditions: a control empty bottle, isotonic saline, silicone oil and water. Each stimulus was presented for 30 s (shaded box). Quantification of the responses is shown in the bottom panel. Activity change (left, area under curve) and baseline activity shift (right,  $\Delta F$  change) were quantified for SFO<sup>nNOS</sup> neurons (GCaMP6s, dark blue bars; control, light blue bars). A significant shift in the baseline activity ( $\Delta F$  change) was observed only in response to water ingestion ( $n = 6$  mice for saline,  $n = 7$  mice for empty, silicone oil and water,  $n = 5$  mice for eYFP). **b**, Shown are representative responses of SFO<sup>nNOS</sup> neurons (blue traces) to an empty bottle, peanut butter, and 300 mM sucrose solution under food-restricted conditions ( $n = 7$  mice for empty and peanut butter,  $n = 5$  mice for sucrose,  $n = 5$  mice for all eYFP recordings). **c**, Activity change per lick was quantified for MnPO<sup>GLP1R</sup> neurons (GCaMP6s, red bars; eYFP, grey bars) under water-restricted conditions (left,  $n = 6$  mice for saline and silicone oil,

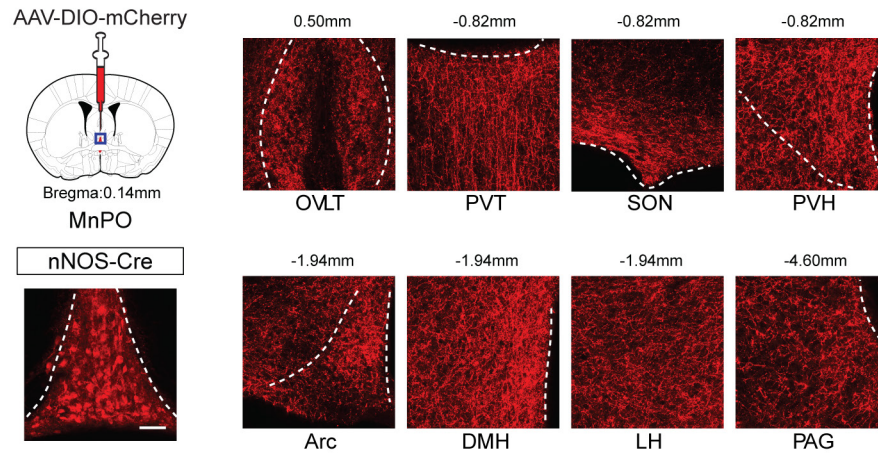
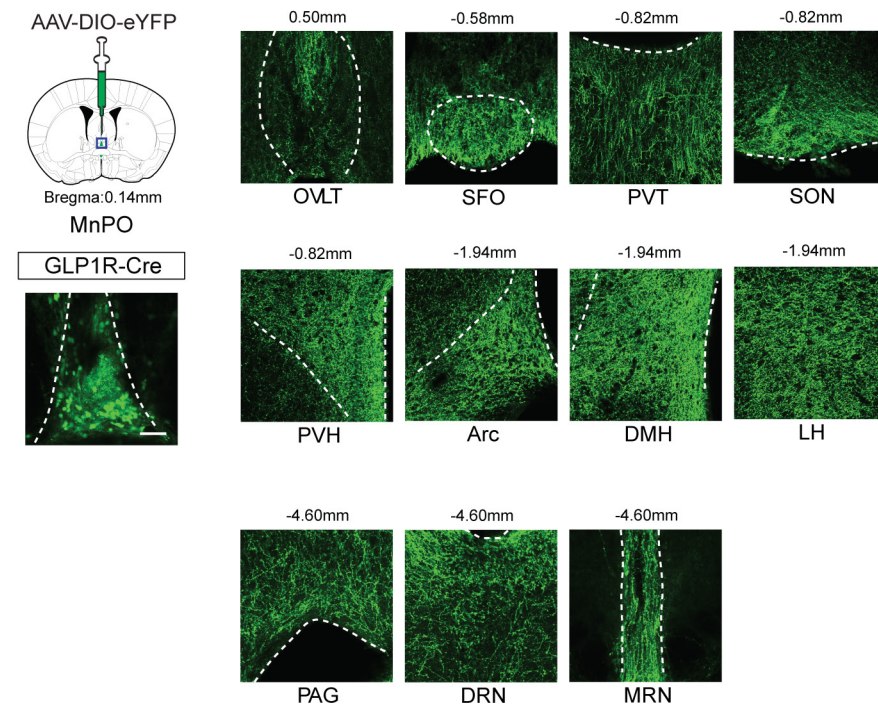
$n = 7$  mice for empty and water,  $n = 6$  mice for all eYFP controls) and food-restricted conditions (right,  $n = 6$  mice for empty and peanut butter,  $n = 7$  mice for sucrose,  $n = 6$  mice for all eYFP controls). All data were reanalysed from Fig. 3b, c. **d**, Normalized fluorescence change of SFO<sup>nNOS</sup> (top) and MnPO<sup>GLP1R</sup> (bottom) neurons from individual mice during licking an empty bottle and water under water-restricted, or sucrose under food-restricted conditions. **e**, MnPO<sup>GLP1R</sup> activation is independent of instinctive need. Left, fibre photometry recording of MnPO<sup>GLP1R</sup> neurons while activating the SFO<sup>nNOS</sup> neurons. GCaMP6s was virally expressed in MnPO<sup>GLP1R</sup> neurons for recording calcium dynamics while activating SFO<sup>nNOS</sup> neurons by hM3Dq-mCherry under the CamKII promoter. Middle, intraperitoneal CNO injection and water deprivation induce water drinking, which robustly activates MnPO<sup>GLP1R</sup> neurons (red and blue traces respectively). Right, activity change (area under the curve) and licks were quantified for natural thirst and CNO activation ( $n = 5$  mice). \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , paired two-tailed  $t$ -test or Kruskal–Wallis one-way ANOVA test with Dunn's correction for multiple comparisons. All error bars show mean  $\pm$  s.e.m.





**Extended Data Figure 8 | Acute inhibition or chronic ablation of MnPO<sup>GLP1R</sup> neurons causes overdrinking.** **a, b**, Acute inhibition of hM4Di-expressing MnPO<sup>GLP1R</sup> neurons by CNO modestly increases water consumption at the onset of drinking. Drinking behaviour was monitored for 30 min after the injection of CNO (**a**); magnified data (0–1 min) is shown in **b** ( $n = 8$  mice). **c, d**, mCherry controls for acute inhibition of MnPO<sup>GLP1R</sup> neurons. Drinking behaviour was monitored for 30 min after the injection of CNO or vehicle under water-deprived conditions with free access to saline (**c**) or water (**d**). No significant difference was found between mice injected with CNO and vehicle ( $n = 6$  mice). **e**, Schematic for the genetic ablation of MnPO<sup>GLP1R</sup> neurons with AAV-flex-Casp3-

TEVp (left) in *Glp1r-cre/Ai9* mice. Compared to a control animal (right), a Casp3-injected animal displayed almost no GLP1R-expressing neurons in the MnPO (middle, representative image from one out of four mice). In both cases, GLP1R-expressing neurons were labelled using *Glp1r-cre/Ai9* transgenic mice. **f**, Genetic ablation of MnPO<sup>GLP1R</sup> neurons (red trace,  $n = 4$  mice) recapitulates the overdrinking phenotype similar to the acute inhibition by hM4Di (Fig. 5b), compared to control eYFP group (black trace,  $n = 6$  mice).  $**P < 0.01$ , by two-tailed Mann–Whitney *U* test. All error bars and shaded areas show mean  $\pm$  s.e.m. Scale bar, 50  $\mu$ m. The mouse brain in this figure has been reproduced from the mouse brain atlas<sup>45</sup>.

**a****b**

**Extended Data Figure 9 | Neural projections from  $nNOS^+$  and  $GLP1R^+$  MnPO neurons. a, b**, Left, schematics for mapping downstream targets of MnPO neurons using AAV-DIO-mCherry (a) or AAV-DIO-eYFP (b). Right, the major outputs from MnPO neurons. *nNOS-cre* (a) and *Glpr-cre* (b) mice were injected with AAV-DIO-mCherry and AAV-DIO-eYFP in the MnPO respectively, and the axon projections were examined using reporter expression. Shown are the injection sites and main representative

downstream targets (one out of three mice). Arc, Arcuate Nucleus; DMH, dorsomedial hypothalamic nucleus; DRN, dorsal raphe nucleus; LH, lateral hypothalamus; MRN, median raphe nucleus; PAG, periaqueductal gray; PVH, paraventricular hypothalamic nucleus; PVT, paraventricular thalamic nucleus; SON, supraoptic nucleus. Scale bars, 50  $\mu$ m. The mouse brain in this figure has been reproduced from the mouse brain atlas<sup>45</sup>.

# Environment dominates over host genetics in shaping human gut microbiota

Daphna Rothschild<sup>1,2\*</sup>, Omer Weissbrod<sup>1,2\*</sup>, Elad Barkan<sup>1,2\*</sup>, Alexander Kurilshikov<sup>3</sup>, Tal Korem<sup>1,2</sup>, David Zeevi<sup>1,2</sup>, Paul I. Costea<sup>1,2</sup>, Anastasia Godneva<sup>1,2</sup>, Iris N. Kalka<sup>1,2</sup>, Noam Bar<sup>1,2</sup>, Smadar Shilo<sup>1,2</sup>, Dar Lador<sup>1,2</sup>, Arnau Vich Vila<sup>3,4</sup>, Niv Zmora<sup>5,6,7</sup>, Meirav Pevsner-Fischer<sup>5</sup>, David Israeli<sup>8</sup>, Noa Kosower<sup>1,2</sup>, Gal Malka<sup>1,2</sup>, Bat Chen Wolf<sup>1,2</sup>, Tali Avnit-Sagi<sup>1,2</sup>, Maya Lotan-Pompan<sup>1,2</sup>, Adina Weinberger<sup>1,2</sup>, Zamir Halpern<sup>7,9</sup>, Shai Carmi<sup>10</sup>, Jingyuan Fu<sup>3,11</sup>, Cisca Wijmenga<sup>3,12</sup>, Alexandra Zhernakova<sup>3</sup>, Eran Elinav<sup>5§</sup> & Eran Segal<sup>1,2§</sup>

**Human gut microbiome composition is shaped by multiple factors but the relative contribution of host genetics remains elusive. Here we examine genotype and microbiome data from 1,046 healthy individuals with several distinct ancestral origins who share a relatively common environment, and demonstrate that the gut microbiome is not significantly associated with genetic ancestry, and that host genetics have a minor role in determining microbiome composition. We show that, by contrast, there are significant similarities in the compositions of the microbiomes of genetically unrelated individuals who share a household, and that over 20% of the inter-person microbiome variability is associated with factors related to diet, drugs and anthropometric measurements. We further demonstrate that microbiome data significantly improve the prediction accuracy for many human traits, such as glucose and obesity measures, compared to models that use only host genetic and environmental data. These results suggest that microbiome alterations aimed at improving clinical outcomes may be carried out across diverse genetic backgrounds.**

The gut microbiome is increasingly recognized as having fundamental roles in human physiology and health<sup>1,2</sup>. A central question is the extent to which microbiome composition is determined by host genetics. Previous studies have identified several heritable bacterial taxa<sup>3–7</sup> but the combined bacterial abundance accounted for by them has not yet been quantified. Other studies have found associations between host single nucleotide polymorphisms (SNPs) and individual bacterial taxa or pathways<sup>5,8–11</sup>. However, most previously reported associations are not statistically significant after multiple testing correction<sup>3</sup>. A recent study identified 42 SNPs that together explained 10% of the variance of microbiome  $\beta$ -diversity<sup>9</sup>. However, the statistical significance of this result has not yet been evaluated. Thus, the extent to which human genetics shape microbiome composition remains unclear.

Here we studied microbial–genetic associations using a cohort of 1,046 healthy Israeli individuals with metagenome-sequenced and 16S rRNA gene-sequenced gut microbiomes, genotypes, anthropometric and blood measurements, and dietary habits<sup>12</sup>. Individuals in our cohort are of several different ancestral origins but we assume, owing to their broadly similar lifestyles, that they share a relatively homogeneous environment.

Our results demonstrate that gut microbiome composition is shaped predominantly by environmental factors. Specifically, we show that the microbiome is not significantly associated with genetic ancestry or with individual SNPs, and that previously reported associations are not replicated across different studies. We further estimate that the average heritability of gut microbiome taxa is only 1.9%, by analysing

data from 2,252 twins from the TwinsUK cohort<sup>6</sup>. However, further and larger-scale studies are required to accurately quantify gut microbiome heritability.

To provide direct evidence that the microbiome is shaped largely by environmental factors, we show that there is significant similarity among the microbiomes of genetically unrelated individuals who share a household, but no significant microbiome similarity among relatives who do not have a history of household sharing. We further demonstrate that over 20% of the variance in microbiome  $\beta$ -diversity can be inferred from environmental factors associated with diet and lifestyle, consistent with previous studies<sup>13,14</sup>.

Because our findings suggest that gut microbiome and host genetics are largely independent, we compare the power of the gut microbiome and of host genetics to predict host phenotypes. We define the term ‘microbiome-association index’ ( $b^2$ ) that—by analogy with genetic heritability, which is typically termed  $h^2$ —quantifies the overall association between the microbiome and a host phenotype after accounting for host genetics. We find significant  $b^2$  levels of 22–36% for body mass index (BMI; 25%), fasting glucose levels (22%), glycaemic status (25%), levels of high-density lipoprotein (HDL) cholesterol (36%), waist circumference (29%), hip circumference (27%), waist–hip ratio (WHR; 24%) and lactose consumption (36%). We note that  $b^2$  should be interpreted with caution, because it is a correlative measure that does not necessarily indicate causality and it may be confounded by environmental factors. We additionally demonstrate that using microbiome data together with human genetic data substantially improves the

<sup>1</sup>Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel. <sup>2</sup>Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 7610001, Israel. <sup>3</sup>University of Groningen, University Medical Center Groningen, Department of Genetics, 9713 GZ Groningen, The Netherlands. <sup>4</sup>University of Groningen, University Medical Center Groningen, Department of Gastroenterology and Hepatology, 9713 GZ Groningen, The Netherlands. <sup>5</sup>Immunology Department, Weizmann Institute of Science, Rehovot 7610001, Israel. <sup>6</sup>Internal Medicine Department, Tel Aviv Sourasky Medical Center, Tel Aviv 6423906, Israel. <sup>7</sup>Research Center for Digestive Tract and Liver Diseases, Tel Aviv Sourasky Medical Center, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 6423906, Israel. <sup>8</sup>Day Care Unit and the Laboratory of Imaging and Brain Stimulation, Kfar Shaul Hospital, Jerusalem Center for Mental Health, Jerusalem 9106000, Israel. <sup>9</sup>Digestive Center, Tel Aviv Sourasky Medical Center, Tel Aviv 6423906, Israel. <sup>10</sup>Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem 9112001, Israel. <sup>11</sup>University of Groningen, University Medical Center Groningen, Department of Pediatrics, 9713 GZ Groningen, The Netherlands. <sup>12</sup>Department of Immunology, K.G. Jebsen Coeliac Disease Research Centre, University of Oslo, 0424 Oslo, Norway.

\*These authors contributed equally to this work.

§These authors jointly supervised this work.



accuracy with which human phenotypes can be predicted, consistent with a previous smaller-scale study<sup>15</sup>.

Finally, we successfully replicate our results in 836 Dutch individuals, with genotypes and metagenomic data, from the LifeLines DEEP (LLD) cohort<sup>8</sup>. Taken together, our results demonstrate that the gut microbiome is predominantly shaped by environmental factors, and is strongly correlated with many human phenotypes after accounting for host genetics.

## Results

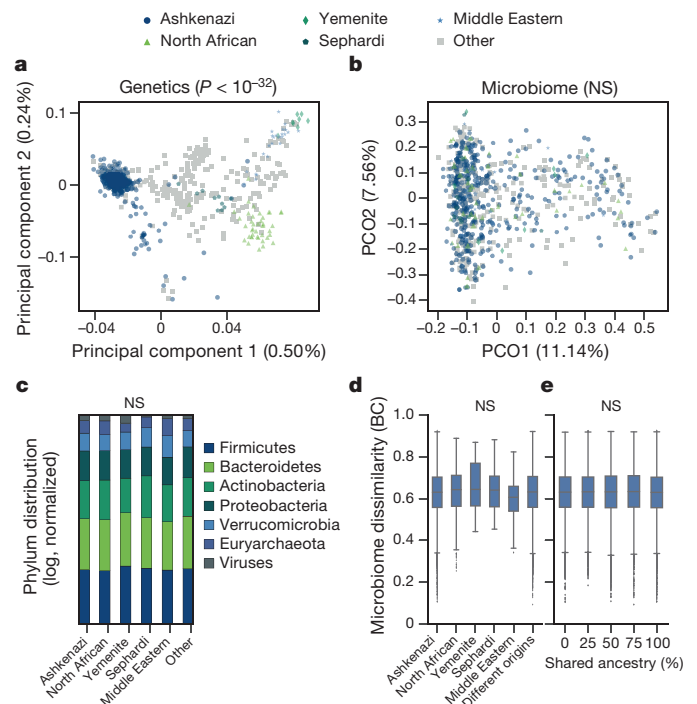
We studied a cohort of 1,046 healthy Israeli adults from whom we collected blood for genotyping and phenotyping, stools for metagenome sequencing and 16S rRNA gene sequencing, anthropometric measurements, and answers to food frequency and lifestyle questionnaires<sup>12</sup> (Extended Data Table 1 and Supplementary Table 1). We performed genotyping at 712,540 SNPs and imputed them to 5,567,647 SNPs (Methods). Stool samples were profiled using both metagenome sequencing and 16S rRNA gene sequencing, and then analysed at multiple taxonomic levels; the results presented here are based on metagenome species analysis (results at metagenome phylum, class, order, family, genus or bacterial gene levels, and for 16S genus and operational taxonomic unit levels, are provided in Supplementary Tables where appropriate). We included covariates for age, gender, stool collection method, and self-reported daily median caloric, fat, protein and carbohydrate consumption (Methods).

### Limited evidence of microbiome–genetic associations

Our sample consists of self-reported Ashkenazi ( $n = 508$ ), North African ( $n = 64$ ), Middle Eastern ( $n = 34$ ), Sephardi ( $n = 19$ ), Yemenite ( $n = 13$ ) and ‘admixed/other’ ( $n = 408$ ) ancestries<sup>16</sup> (Supplementary Table 2). We first successfully verified that the top two host genetic principal components are strongly associated with self-reported ancestry ( $P < 10^{-32}$  for both principal component 1 and principal component 2, Kruskal–Wallis test; Fig. 1a, Extended Data Table 2 and Supplementary Table 3). By contrast, we found no significant association between ancestry and microbiome composition. Specifically, there was no significant correlation between any of the top five host genetic principal components and any of the top five microbiome  $\beta$ -diversity principal coordinates (PCOs, computed using Bray–Curtis dissimilarity;  $P > 0.49$  for all pairwise associations, Spearman correlation; Supplementary Table 4).

We also found no significant differences between ancestries in terms of microbiome composition (quantified by PCOs of Bray–Curtis dissimilarities),  $\alpha$ -diversity (quantified by the Shannon diversity index) or abundance of specific taxa (Kruskal–Wallis test for non-admixed individuals; Fig. 1b–d, Extended Data Table 2 and Supplementary Table 3). We obtained similar results when testing whether individuals who are more ancestrally similar, quantified by the fraction of grandparents from the same ancestry, have more similar microbiomes (quantified by Bray–Curtis dissimilarity),  $\alpha$ -diversity or abundance of specific taxa (Mantel test<sup>17</sup>; Methods, Fig. 1e, Extended Data Table 2 and Supplementary Table 5). As a control, we verified that ancestrally similar individuals are significantly similar in terms of their genetics ( $P < 10^{-5}$ , Mantel test; Methods, Extended Data Table 2 and Supplementary Table 5). We also found no significant association between microbiome composition and genetic kinship (Fig. 2a, Extended Data Table 2 and Supplementary Table 6), though we note that SNP-based kinship tests are less powerful than ancestry-based tests (Supplementary Tables 7–10 and Supplementary Information).

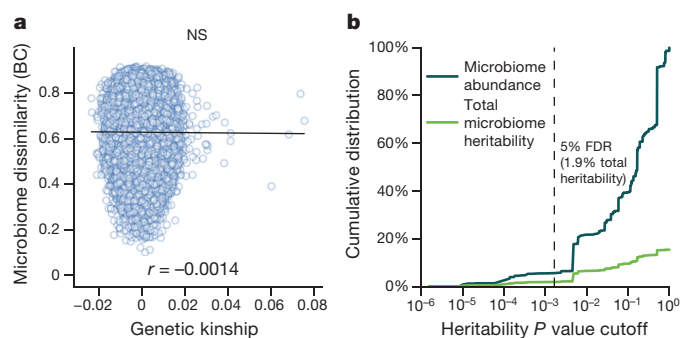
One caveat of our study is the presence of imbalanced per-population sample sizes. Although an ideal study should have equal per-population sample sizes, we verified that our study is well-powered to detect microbiome–ancestry associations. Specifically, we found that the probability of finding a statistically significant microbiome–ancestry association is 70% if only 10% of the microbiome variance is explained



**Figure 1 | Genetic ancestry is not significantly associated with microbiome composition.** **a**, Genetic principal components are strongly associated with self-reported ancestry, with Ashkenazi ( $n = 345$ ), North African ( $n = 42$ ), Middle Eastern ( $n = 24$ ), Sephardi ( $n = 10$ ), Yemenite ( $n = 8$ ) and admixed/other (other) ( $n = 286$ ) ancestries ( $P < 10^{-32}$ ; Kruskal–Wallis). **b**, As in **a**, but for microbiome principal coordinate analysis ( $P > 0.08$ ; Kruskal–Wallis). **c**, The distribution of average phylum abundance among 582 non-admixed individuals (in log scale, normalized to sum to 1.0) is not associated with ancestry ( $P > 0.05$ ; Kruskal–Wallis). NS, not significant. **d**, Box plots of Bray–Curtis (BC) dissimilarities across all pairs of 737 individuals for whom the ancestries of all grandparents are known, demonstrating that microbiome composition is not associated with ancestry ( $P > 0.06$ ; Kruskal–Wallis test for the top five Bray–Curtis PCOs).  $n = 105,570$  (Ashkenazi), 1,711 (North African), 528 (Middle Eastern), 136 (Sephardi) and 78 (Yemenite) same ancestry pairs;  $n = 61,048$  different ancestry pairs. The lower and upper limits of the boxes represent the 25% and 75% percentiles, respectively, and the top and bottom whiskers represent the 5% and 95% percentiles, respectively. **e**, Box plots of Bray–Curtis dissimilarities across pairs of 946 individuals (including admixed individuals), organized according to shared ancestry fraction (the fraction of grandparents of the same ancestry), for pairs with 0% ( $n = 167,618$ ), 25% ( $n = 33,119$ ), 50% ( $n = 100,163$ ), 75% ( $n = 34,187$ ) and 100% ( $n = 111,898$ ) shared ancestry fractions. The lower and upper limits of the boxes represent the 25% and 75% percentiles, respectively, and the top and bottom whiskers represent the 5% and 95% percentiles, respectively. The figure demonstrates that microbiome similarity is not associated with ancestral similarity ( $P = 0.73$ ; Mantel test).

by ancestry, and is greater than 90% if over 30% of the microbiome variance is explained by ancestry (Supplementary Information).

The lack of association between microbiome composition and genetic ancestry suggests that the microbiome is not strongly associated with host genetics. Because twin studies are ideal for heritability estimation<sup>18</sup>, we analysed a previously studied<sup>6</sup> dataset of 2,252 twins to directly quantify microbiome heritability. First, we found that the sum of the relative abundances of all 33 taxa reported as significantly heritable in the previous study<sup>6</sup> accounted for only 5.6% of total microbiome composition (Methods). Next, we estimated the overall microbiome heritability using the formula  $H^2 = \sum_{t \in S} r_t h_t^2$ , in which  $r_t$  and  $h_t^2$  are the relative abundance and estimated heritability of taxon  $t$ , respectively, and  $S$  is the set of significantly heritable taxa (making sure not to count the same taxon multiple times; see Methods). The resulting heritability estimate was only 1.9% or, at most, 8.1% when performing



**Figure 2 | Genetic kinship is weakly associated with microbiome composition.** **a**, Scatter plot of genetic kinship of pairs of individuals ( $x$  axis) and their microbiome dissimilarity ( $y$  axis), among all pairs of  $n = 715$  unrelated genotyped individuals, demonstrating that genetic kinship and microbiome similarity are uncorrelated ( $P = 0.59$ ; Mantel test). NS, not significant. **b**, The overall heritability of significantly heritable taxa in a cohort of 2,252 twins<sup>6</sup> (light green) and their cumulative relative abundance (dark green). The  $x$  axis indicates the  $P$ -value cutoff required to declare a taxon as significantly heritable (using  $P$  values computed in a previous study<sup>6</sup>). The figure demonstrates that the overall microbiome heritability is small regardless of the cutoff. Under a cutoff corresponding to a 5% FDR, the overall microbiome heritability is 1.9%. FDR, false discovery rate.

no correction for multiple testing in the definition of  $S$  (Fig. 2b and Supplementary Table 11). These numbers serve as estimates of the lower and upper bound of the true overall microbiome heritability.

In addition, we applied several machine-learning algorithms to predict ancestry proportions from microbiomes, but none were successful (prediction  $R^2 < 0.01$  for all ancestries; Methods). We also tried to predict top microbiome PCOs from ancestral or genetic data, and again found no significant associations ( $P > 0.1$ , permutation testing; Methods and Supplementary Table 12).

Finally, we verified that similar results are obtained when repeating the above experiments using any of the following: other metagenome-derived taxonomic and functional levels (phylum, class, order, family, genus and bacterial genes; see Methods); 16S rRNA gene sequencing; Unifrac- and Jaccard-based dissimilarity measures; non-metric multidimensional scaling<sup>17</sup> instead of principal coordinate analysis; dichotomization of relative abundance into presence/absence patterns; and when omitting covariates (Extended Data Fig. 1 and Supplementary Tables 3–6, 12).

We next investigated associations between individual SNPs and microbiome  $\beta$ -diversity, using a distance-based  $F$  test<sup>19</sup> (Extended Data Fig. 2 and Methods). This analysis found two loci with marginal genome-wide significance (rs6563994,  $P = 3 \times 10^{-8}$ ; rs13149273,  $P = 4.2 \times 10^{-8}$ ). However, as we show later, we could not replicate these findings in an additional cohort. In addition, our data did not replicate any of the 42 SNPs previously reported<sup>9</sup> as being significantly associated with microbiome  $\beta$ -diversity, either when using an  $F$  test or the previously applied method<sup>9</sup> ( $P > 0.05$  for all previously reported SNPs; Methods).

The previous study<sup>9</sup> reported that these 42 SNPs could be used to infer 10% of the  $\beta$ -diversity variance, but did not report the statistical significance of this result. We were able to explain 12.1% of  $\beta$ -diversity variance using the 42 SNPs that were most closely associated with  $\beta$ -diversity in our own data, but this result was not statistically significant under permutations ( $P = 0.74$ ; Methods). We conclude that inferring  $>10\%$  of  $\beta$ -diversity variance using top-ranked SNPs may be an inherent property of the method used rather than a biologically meaningful result. Thus, we find very limited evidence in our data for the association of any individual SNP with microbiome  $\beta$ -diversity.

We next tested for associations between individual SNPs and specific taxa, using a linear mixed model (LMM) and dichotomization of zero-inflated taxa (Methods; Supplementary Table 13). This analysis identified 43 loci with  $P < 5 \times 10^{-8}$ , but none remained

statistically significant at a false discovery rate (FDR) of 5% (Fig. 3a and Supplementary Table 14).

We also investigated the association of 225 SNPs in 211 loci reported as significantly associated with specific taxa or with  $\beta$ -diversity in any of five previous studies<sup>5,6,8–10</sup> (Methods). To maximize replication power, we used the minimal  $P$  value obtained for each SNP across all taxa belonging to the same phylum. Only 7 of the 211 loci (3.3%) replicated at  $P < 0.05/211$  (Fig. 3b and Supplementary Table 15; Methods). Two of these seven loci are found close to the *LCT* gene, which encodes the lactase enzyme that enables lactose consumption, and were found by previous studies<sup>6–10</sup> to be associated with *Bifidobacterium*, possibly owing to its association with lactose consumption.

Notably, the *LCT* gene is the only case in which there was an overlap between the SNPs reported in any pair of the five previous studies<sup>5,6,8–10</sup>. Moreover, no pair of previously reported SNPs from any two studies were within 100 kb of one another, or within 1 Mb of one another and associated with at least one bacterial taxa of the same phylum (Supplementary Table 15).

### Microbiome–environment associations

We next investigated whether 24 pairs of related individuals—using second- to fifth-degree relatives—with no history of household sharing had a similar microbiome composition, when compared to non-related pairs with no household sharing (using Bray–Curtis dissimilarity). We found no such evidence of similar microbiomes ( $P > 0.4$ , permutation testing; Methods; Extended Data Fig. 3, Supplementary Table 16). By contrast, when investigating 55 first-degree-relative pairs, who are likely to have a history of household sharing, we found significant similarities in their microbiomes at the genus and species taxonomic levels, and at the level of bacterial genes ( $P < 5 \times 10^{-3}$ ; Methods; Extended Data Fig. 3 and Supplementary Table 16).

To test the effect of recent household sharing, we repeated the above analysis for 32 pairs of genetically unrelated individuals who reported sharing a household, and again found significant microbiome similarities at the level of species and of bacterial genes ( $P < 2 \times 10^{-3}$ ; Extended Data Fig. 3 and Supplementary Table 16).

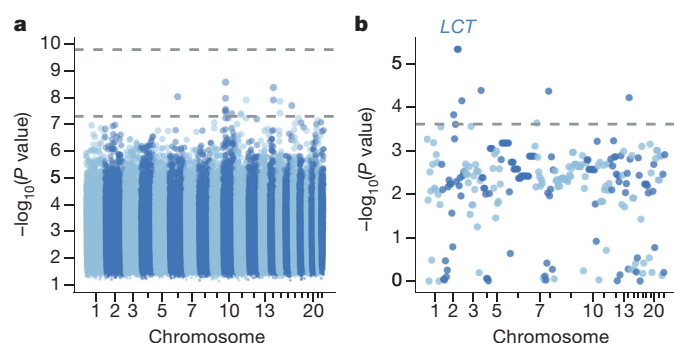
These results suggest that past or present household sharing may partly determine gut microbiome composition, whereas we find very little supporting evidence for microbiome similarities among relatives with no past household sharing. Our results corroborate previous studies<sup>20</sup>, including a recent twin study that showed that the gut microbiomes of twins become more genetically dissimilar over time when living apart<sup>11</sup>, and another study that showed that microbiome similarity among monozygotic twins compared to dizygotic twins is only marginally significant<sup>4</sup> ( $P = 0.032$  under an unweighted UniFrac dissimilarity,  $P > 0.05$  under Bray–Curtis and weighted UniFrac dissimilarity).

We next directly quantified the fraction of microbiome  $\beta$ -diversity variance that can be inferred from environmental factors (based on Bray–Curtis dissimilarities) using answers to food frequency, lifestyle and drug use questions, as well as blood measurements, self-reported median daily intake of calories, protein, fat and carbohydrates, age, gender, glycaemic status, BMI, fasting glucose levels and the top 5 host genetic principal components (Methods; Supplementary Table 17).

We used a feature-selection algorithm that selected 95 environmental features that together could be used to infer 20.03% of the variance of microbiome  $\beta$ -diversity via PERMANOVA<sup>21</sup> (Methods,  $P < 0.002$ ; Extended Data Fig. 4 and Supplementary Table 18), consistent with previous studies<sup>13,14</sup>. By contrast, host SNPs could not be used to infer a statistically significant fraction of  $\beta$ -diversity variance ( $P = 0.11$ , Methods).

### Microbiome–phenotype associations

We next investigated how well host phenotypes can be inferred on the basis of the microbiome as compared to host genetics. The fraction of phenotypic variance that can be inferred from the microbiome after accounting for other explanatory variables including host genetics (that



**Figure 3 | Limited evidence for microbiome associations with specific SNPs.** **a**, Manhattan plot showing the lowest  $P$  value obtained at every SNP tested for association with 313 taxa (computed by FaST-LMM<sup>38</sup> using  $n = 814$  individuals) and with microbiome  $\beta$ -diversity (computed with a distance-based  $F$  test using  $n = 715$  non-related individuals). The dashed lines represent a genome-wide significant  $P$  value corrected (top line) and not corrected (bottom line) for testing 313 taxa. **b**, The lowest  $P$  value obtained across 313 taxon association tests for each of 225 SNPs in 211 loci previously reported to be significantly associated with the microbiome<sup>5,6,8–10</sup> (computed by FaST-LMM using  $n = 814$  individuals). Seven SNPs are successfully replicated at  $P < 0.05/211$  (dashed line; rs4988235, rs6730157, rs7656342, rs10112815, rs11626933, rs56006724 and rs7782745), two of which reside near the *LCT* gene.

is,  $b^2$ ), represents a formal measure of predictability: larger  $b^2$  values indicate that the microbiome is more informative with respect to a phenotype of interest. We estimate the value of  $b^2$  in a narrow sense that cannot capture gene–gene or gene–environment interactions, as is common in heritability estimation<sup>22</sup> (Supplementary Information).

Heritability estimation is typically performed in an LMM framework<sup>23</sup> and requires a kinship matrix, which is typically estimated from SNPs<sup>24</sup>. We define the analogous bacterial kinship matrix on the basis of bacterial genes (Methods). We note, however, that the results may be confounded by unmeasured environmental factors, and that  $b^2$  cannot be used to determine causality because microbiome composition can both affect and be affected by host phenotypes.

We used FIESTA<sup>25</sup> to verify that  $b^2$  is consistently estimated in a more reliable fashion than  $h^2$  for a given sample size, and that samples as large as 4,000 individuals are required for  $h^2$  estimates to be as accurate as the  $b^2$  estimate obtained in our cohort of 715 unrelated genotyped individuals (based on previous genetic data<sup>26</sup>, Methods; Fig. 4a and Supplementary Table 19). We conclude that  $b^2$  estimation is more accurate than  $h^2$  estimation for a given sample size, and can be carried out with hundreds rather than thousands of individuals.

We next estimated  $b^2$  for several phenotypes of interest (Extended Data Table 1), and used polygenic risk scores (PRS) based on summary statistics as an additional covariate to account for host genetic factors (Methods; Supplementary Table 20). We found 8 of the 12 traits we investigated to be significantly associated with the microbiome, after accounting for age, gender, diet and host genetics, with estimated  $b^2$  levels of 36% for non-fasting HDL cholesterol levels, 36% for lactose consumption, 29% for waist circumference, 27% for hip circumference, 25% for BMI, 25% for glycaemic status, 24% for WHR and 22% for fasting glucose (Methods; Fig. 4b, c, Extended Data Fig. 5a, d, g, j and Supplementary Tables 21, 22). These  $b^2$  estimates are comparable to previous SNP heritability estimates<sup>27–34</sup> (Fig. 4b and Supplementary Table 21), which indicates that the microbiome is strongly associated with these traits.

To provide another comparison between host genetics and microbiome, we evaluated the ability of a linear prediction model (Methods) to predict human phenotypes from bacterial gene abundances, PRS, age, sex, as well as daily median caloric, carbohydrate, fat and protein consumption. The contribution of a specific data source to the phenotype can be assessed by the reduction in prediction power when

excluding this data source. We found that the prediction accuracy for 10 of the 12 traits we investigated—including BMI, HDL cholesterol and fasting glucose levels—is substantially improved when microbiome data is added to PRS (Fig. 4d, Extended Data Fig. 5b, c, e, f, h, i, k, l and Supplementary Table 23). Moreover, the contribution of both data sources is largely additive, consistent with our finding that microbiome and host genetics are largely independent of one another (Fig. 4e).

Taken together, these results demonstrate that host genetics and microbiome are complementary for predicting host phenotypes, and that phenotype prediction can be substantially improved by using both host genetics and microbiome data.

### Independent validation on the LLD cohort

To verify that our results are broadly applicable across different study designs and populations, we repeated our analyses on a cohort of 836 Dutch individuals from the LLD cohort<sup>8</sup>, with metagenome-sequenced gut microbiomes, genotypes and the same covariates that we used in our analysis of the Israeli cohort.

In the LLD cohort, as in the Israeli cohort, there were no statistically significant associations between top host genetic principal components and microbiome PCOs (Supplementary Table 24), or between genetic kinship and microbiome  $\beta$ -diversity or bacterial taxa (Supplementary Table 25). A meta-analysis of both studies did not yield significant associations (Methods). Notably, the combined dataset is the largest cohort (1,551 individuals) of genotypes and metagenomic sequenced gut microbiomes analysed to date.

Next, we searched for significant associations between SNPs and microbiome  $\beta$ -diversity in the LLD cohort via the distance-based  $F$  test<sup>19</sup> but could not replicate the two marginally-significant loci found in the Israeli cohort. Two loci had a genome-wide significance ( $P < 5 \times 10^{-8}$ ) in the LLD cohort (Supplementary Table 26). One of these SNPs (rs4988235) is associated with lactase persistence, was replicated in the Israeli cohort ( $P = 0.018$ ) and has previously been reported in microbiome–genetic association studies<sup>6–8,10</sup>. This was also the only genome-wide significant SNP in a meta-analysis of the Israeli and LLD cohorts ( $P = 2.2 \times 10^{-9}$ ), and is the only SNP reported by multiple studies as being associated with the gut microbiome. Thus, multiple lines of evidence suggest that rs4988235 is the SNP most strongly associated with gut microbiome composition.

Next, we found that all phenotypes that were significantly associated with the microbiome in the Israeli cohort, according to the  $b^2$  measure, were also significantly associated in the LLD cohort, with the exception of WHR (Extended Data Fig. 5m and Supplementary Table 27). We performed phenotype predictions using the LLD cohort, and observed results highly similar to those obtained when using data from the Israeli cohort (Extended Data Fig. 5n, o and Supplementary Table 28).

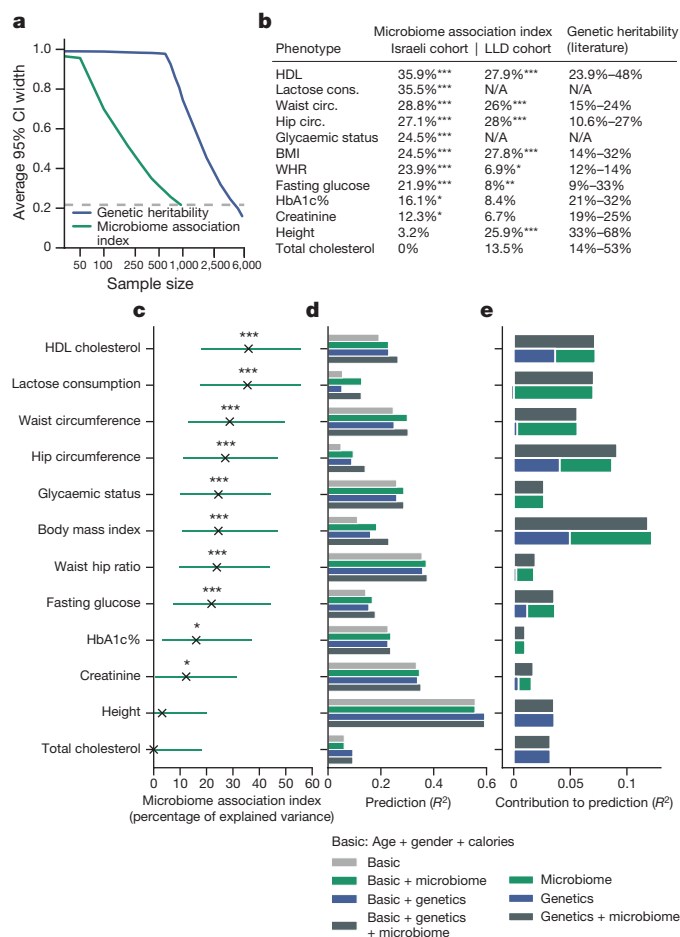
Overall, we conclude that there is considerable congruency between the results obtained in the two cohorts, despite the substantial differences in the study populations, data collection method and processing pipeline. This demonstrates that our results are robust to diverse populations and experimental settings.

### Discussion

In this study we used a range of statistical analyses across multiple cohorts, all of which led to the conclusion that the environment has a substantially greater role than host genetics in shaping the human gut microbiome. Several recent studies have reported that the microbiome is not only stable over time<sup>35,36</sup> but also—to some extent—resilient to perturbations such as antibiotics and pathogens<sup>37</sup>; the extent and determinants of such stability remain unresolved. As a small minority of heritable microbes are unlikely to generate this stability, it will be interesting to discover which mechanisms underlie microbiome stability and which perturbations cause the dysbiosis that can lead to disease susceptibility.

We proposed  $b^2$  as a means of quantifying microbiome association with host phenotypes, and showed that  $b^2$  can be reliably estimated





**Figure 4 | The gut microbiome can be used to infer a significant fraction of the variance of several human phenotypes.** **a**, For a given sample size,  $b^2$  can be estimated more accurately than genetic heritability, as evaluated by using up to 946 individuals with gut microbiomes and up to 5,652 genotyped individuals from the Wellcome Trust Case Control Consortium 2 (ref. 26). Smaller 95% confidence interval (CI) widths indicate a greater confidence in the estimation. **b**,  $b^2$  estimates from the analysis of 715 individuals with measured genotyped and gut microbiomes from the Israeli cohort (left column) and of 836 individuals from the LLD cohort (middle column) are comparable to previous genetic heritability estimates<sup>27–34</sup> (right column). \*FDR < 0.05, \*\*FDR < 0.01 and \*\*\*FDR < 0.001. Cons., consumption, circ., circumference. **c**,  $b^2$  estimates of several human phenotypes and their 95% confidence intervals, evaluated using 715 individuals. \*FDR < 0.05, \*\*FDR < 0.01 and \*\*\*FDR < 0.001. **d**, Phenotype prediction accuracy for 715 individuals, evaluated using a LMM under different sets of predictive features (measured using coefficient of determination ( $R^2$ )), using four different models for each phenotype: (i) ‘Basic’, age, gender and diet features; (ii) ‘Basic + microbiome’, basic features and relative abundances of bacterial genes; (iii) ‘Basic + genetics’, basic features and host genotypes; and (iv) ‘Basic + genetics + microbiome’: basic features, relative abundances of bacterial genes and host genotypes. **e**, The additive contribution of microbiome and genetics to prediction performance evaluated using a LMM across 715 individuals, over a model that includes only basic features. The joint contribution of microbiome and genetics is similar to the sum of the individual contributions, suggesting these are independent contributions.

using metagenomic cohorts of only hundreds of individuals; we then found that several phenotypes exhibit substantial  $b^2$  levels, in the range of 22–36%. Finally, we showed that adding microbiome data to host genetics data improves prediction accuracy for several host phenotypes, and that the two data sources contribute additively. We note that  $b^2$  should be interpreted with caution as it is a correlative measure and may be confounded by environmental factors.

Previous studies have identified heritable bacteria by observing co-occurrence among family members<sup>4–6,11</sup>, or by reporting associations between specific SNPs and bacterial taxa<sup>5,6,8–11</sup>. Our results are consistent with these published data, and collectively suggest that only a small number of bacterial taxa are likely to be strongly heritable, and that most SNP–bacteria associations are either weak or population-dependent. Our re-analysis of a recent study of twins<sup>6</sup> estimates that the overall microbiome heritability lies between 1.9% and 8.1%. Future studies with larger sample sizes will probably identify additional heritable taxa and SNP associations, but are unlikely to change the overall conclusion that microbiome composition is predominantly shaped by non-genetic factors.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 June 2017; accepted 16 January 2018.

Published online 28 February 2018.

- Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
- Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: an integrative view. *Cell* **148**, 1258–1270 (2012).
- Kurilshikov, A., Wijmenga, C., Fu, J. & Zhernakova, A. Host genetics and gut microbiome: challenges and perspectives. *Trends Immunol.* **38**, 633–647 (2017).
- Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
- Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
- Goodrich, J. K. *et al.* Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**, 731–743 (2016).
- Goodrich, J. K., Davenport, E. R., Clark, A. G. & Ley, R. E. The relationship between the human genome and microbiome comes into view. *Annu. Rev. Genet.* **51**, 413–433 (2017).
- Bonder, M. J. *et al.* The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
- Wang, J. *et al.* Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* **48**, 1396–1406 (2016).
- Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).
- Xie, H. *et al.* Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584 (2016).
- Zeevi, D. *et al.* Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
- Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
- Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
- Fu, J. *et al.* The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circ. Res.* **117**, 817–824 (2015).
- Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466**, 238–242 (2010).
- Legendre, P. & Legendre, L. *Numerical Ecology* Vol. 24, 3rd edn (Elsevier, 2012).
- Visscher, P. M. & Goddard, M. E. A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. *Genetics* **199**, 223–232 (2015).
- Rühlemann, M. C. *et al.* Application of the distance-based  $F$  test in an mGWAS investigating  $\beta$  diversity of intestinal microbiota identifies variants in SLC9A8 (NHE8) and 3 other loci. *Gut Microbes* <https://doi.org/10.1080/19490976.2017.1356979> (2017).
- Song, S. J. *et al.* Cohabiting family members share microbiota with one another and with their dogs. *eLife* **2**, e00458 (2013).
- McArdle, B. H. & Anderson, M. J. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**, 290–297 (2001).
- Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
- Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
- Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
- Schweiger, R. *et al.* in *RECOMB 2017: Research in Computational Molecular Biology* (ed. Sahinalp, S.) 241–256 (Springer, 2017).
- Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.* **42**, 985–990 (2010).

27. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
28. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
29. Ge, T., Chen, C. Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* **13**, e1006711 (2017).
30. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* **9**, e1003520 (2013).
31. Vattikuti, S., Guo, J. & Chow, C. C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet.* **8**, e1002637 (2012).
32. Arpegård, J. *et al.* Comparison of heritability of cystatin C- and creatinine-based estimates of kidney function and their relation to heritability of cardiovascular disease. *J. Am. Heart Assoc.* **4**, e001467 (2015).
33. Xia, C. *et al.* Pedigree- and SNP-associated genetics and recent environment are the major contributors to anthropometric and cardiometabolic trait variation. *PLoS Genet.* **12**, e1005804 (2016).
34. Heckerman, D. *et al.* Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl Acad. Sci. USA* **113**, 7377–7382 (2016).
35. Antonopoulos, D. A. *et al.* Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation. *Infect. Immun.* **77**, 2367–2375 (2009).
36. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12**, R50 (2011).
37. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
38. Widmer, C. *et al.* Further improvements to linear mixed models for genome-wide association studies. *Sci. Rep.* **4**, 6874 (2014).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank the Segal and Elinav group members for discussions; J. Goodrich for sharing the processed twins microbiome data with us; and participants and staff of the LifeLines DEEP cohort for their collaboration. S.C. thanks the Abisch–Frenkel Foundation. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from [www.wtccc.org.uk](http://www.wtccc.org.uk). Funding for the project was provided by the Wellcome Trust under awards 076113 and 085475. E.S. is supported by the Crown Human

Genome Center; the Else Kroener Fresenius Foundation; D. L. Schwarz; J. N. Halpern; L. Steinberg; and grants funded by the European Research Council and the Israel Science Foundation. E.E. is supported by Y. and R. Ungar, the Gurwin Family Fund for Scientific Research, the Leona M. and Harry B. Helmsley Charitable Trust, the Israel Science Foundation and the Helmholtz Foundation. E.E. holds the Sir Marc and Lady Tania Feldmann Professorial Chair in Immunology, is a senior fellow of the Canadian Institute for Advanced Research, and is an international scholar at the Bill and Melinda Gates Foundation and Howard Hughes Medical Institute. D.R. received a Levi Eshkol PhD Scholarship for Personalized Medicine by the Israeli Ministry of Science. LLD was made possible by grants from the Top Institute Food and Nutrition (GH001) to C.W. C.W. is funded by a European Research Council (ERC) advanced grant (FP/2007-2013/ERC grant 2012-322698), a Netherlands Organization for Scientific Research (NWO) Spinoza prize (NWO SPI 92-266) and the Stiftelsen Kristian Gerhard Jebsen foundation (Norway). A.Z. holds a Rosalind Franklin Fellowship (University of Groningen), ERC starting grant (715772) and NWO Vidi grant (178.056). J.F. is funded by an NWO Vidi grant (NWO-VIDI 864.13.013). A.Z. and J.F. are also funded by CardioVascuair Onderzoek Nederland (CVON 2012-03).

**Author Contributions** D.R., O.W. and E.B. conceived the project, designed and conducted all analyses, interpreted the results, wrote the manuscript and are listed in random order. A.K., A.V.V., J.F., C.W. and A.Z. performed the analyses of the Dutch cohort and interpreted the results. T.K., D.Z. and A.W. designed protocols and supervised data collection. T.K., D.Z., P.I.C., A.G., I.N.K. and N.B. conducted microbiome analyses. S.S. and D.L. designed nutritional and drug databases. N.Z., M.P.-F., D.I. and Z.H. coordinated and supervised clinical aspects of data collection. N.K., G.M. and B.C.W. coordinated and designed data collection. T.A.-S., M.L.-P. and A.W. developed protocols and performed genotyping and microbiome sequencing. S.C. designed the genetic analyses. E.E. and E.S. conceived and directed the project and analyses, designed the analyses, interpreted the results and wrote the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to E.E. ([eran.elinav@weizmann.ac.il](mailto:eran.elinav@weizmann.ac.il)) or E.S. ([eran.segal@weizmann.ac.il](mailto:eran.segal@weizmann.ac.il)).

**Reviewer Information** *Nature* thanks M. Georges and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

**Description of cohorts.** This study used a previously described<sup>12</sup> cohort of individuals collected in Israel. Study participants were healthy individuals aged between 18 and 70 (for full inclusion and exclusion criteria, see previous description<sup>12</sup>). Before the study, participants answered medical, lifestyle and nutritional questionnaires. All participants were monitored by a continuous glucometer (CGM) for seven days. During this period, participants were instructed to record all daily activities, including standardized and real-life meals, in real-time using their smartphones. All participants were genotyped using Illumina OMNI-EXPRESS array and provided stool samples, which were collected using a swab (88% of the individuals) or an OMNIGENE-GUT (12% of the individuals; OMR-200; DNA Genotek) stool collection kit. The stool samples were metagenome-sequenced using Illumina NextSeq and HiSeq, and 16S rRNA gene sequenced as previously described<sup>12</sup>. We validated that SNPs extracted from human reads in pre-filtered metagenomic sequences matched SNPs extracted from the blood of their human host. We further verified that the stool collection method did not confound our results by repeating all analyses using only stool samples collected via swab, which yielded results nearly identical to those obtained under the full dataset (results not shown).

The replication LifeLines DEEP cohort includes 1,539 individuals (636 males and 903 females, age range 18–84 years) from the north Netherlands. We included 836 participants in the analysis, after excluding related individuals, unhealthy individuals and individuals without genotype data or metagenomics sequencing data.

Genome-wide genotyping for the LLD participants was performed using Illumina HumanCytoSNP-12 and Immunochip arrays, and then imputed using Haplotype Reference Consortium server with HRC 1.0 panel. Metagenomics sequencing was performed using the Illumina HiSeq platform. Within 2 weeks of participants giving a blood sample, they collected faecal samples at home and stored them immediately at  $-20^{\circ}\text{C}$ . After transport to the research laboratory on dry ice, faecal samples were stored at  $-80^{\circ}\text{C}$ . Aliquots were made, and DNA was isolated with the AllPrep DNA/RNA Mini kit (Qiagen, 80204) with the addition of mechanical lysis. Reads were quality-filtered, and adapter removal was performed using Trimmomatic<sup>39</sup> (v.0.32). An average of 3.0 Gb of data (around 32.3 million high-quality reads) was obtained per sample. Reads belonging to the human genome were removed by mapping the data to the human reference genome (version NCBI37) with Bowtie2<sup>40</sup> (v.2.1.0). The profile of microbial composition was determined using MetaPhlAn<sup>41</sup> 2.2.

**Genotypes preprocessing and imputation.** We performed stringent quality control in our initial set of 862 genotyped individuals and 712,540 SNPs. We excluded SNPs with a missingness rate  $>5\%$ , Hardy–Weinberg  $P < 10^{-9}$ , minor allele frequency  $<5\%$ ,  $P < 0.01$  for differential missingness between two batches of individuals, or a logistic regression  $P < 10^{-6}$  for separation of the two batches, yielding 545,325 SNPs for subsequent analyses. We additionally excluded individuals with  $>10\%$  missing SNPs, leaving 833 individuals.

Genotypes were pre-phased using EAGLE2<sup>42</sup> without a reference panel, and imputed using IMPUTE2<sup>43</sup> using the 1000 genomes dataset<sup>44</sup> and 128 Ashkenazi Jewish individuals<sup>45</sup> as reference panels. We retained only SNPs with imputation probability  $>90\%$ , and applied the filtering stages above to the imputed data, yielding 5,567,647 imputed SNPs.

**Microbial preprocessing.** Preprocessing of 16S rRNA gene sequencing was performed as previously described<sup>12</sup>, with the addition of rarefaction to 10,000 reads. Weighted and unweighted UniFrac matrices for 16S rRNA gene sequencing samples were computed using QIIME via the beta\_diversity script<sup>46</sup>.

For metagenome analysis, we filtered metagenomic reads containing Illumina adapters, filtered low-quality reads and trimmed low-quality read edges. We detected host DNA by mapping with GEM<sup>47</sup> to the human genome (hg19) with inclusive parameters, and removed human reads. We subsampled all samples to have at most 10 million reads. Relative abundances from metagenomic sequencing were computed using MetaPhlAn2<sup>41</sup> with default parameters. MetaPhlAn relative abundances were capped at a level of  $10^{-4}$ . We removed individuals with  $<15$  observed species from the analysis. After all genotyping and metagenomics quality-control steps, 946 individuals with metagenomics data remained, 814 of whom were genotyped. Unless stated otherwise, we additionally excluded individuals self-reported to share a household and related individuals (using up to three degrees of relationship) from the analysis, yielding  $n = 715$  genotyped individuals (Supplementary Table 1).

When testing for associations between specific taxa and specific SNPs, we log-transformed the data and used only taxa present in at least 5% of individuals in our cohort, which left 7/18 (remaining/total) phyla, 13/28 classes, 17/43 orders, 35/96 families, 80/221 genera and 184/652 species.

**Gene mapping.** We performed gene mapping for the gene-based analyses by computing the length-normalized relative abundances of genes, obtained by similarity mapping with GEM to the gene reference catalogue<sup>48</sup> followed by abundance correction using an iterative algorithm based on Pathoscope<sup>49</sup>, and normalization to sum to 1.0, using single-end reads.

**Fasting glucose phenotyping.** In the  $b^2$  and phenotype prediction analyses, the fasting glucose phenotype was taken from data recorded by CGMs over the course of a week, as previously described<sup>12</sup>. The median glucose measurement over a period of 30 min from self-reported wake-up time was used as a surrogate measure for fasting glucose.

**Glycaemic status.** For each patient we computed a quantity which we term glycaemic status<sup>5</sup> that can serve as an indicator of hyperglycaemia, based on HbA1c, fasting glucose, response to standardized meals<sup>12</sup>, and top glucose percentiles and glucose noise as obtained from the CGM over the course of one week. Each individual was first ranked according to each feature. The glycaemic status of each individual was defined as the median of the ranks of (i) HbA1c; (ii) fasting glucose; (iii) median response to standardized meals; (iv) median of 90%, 95% and 98% glucose percentiles; and (v) glucose noise. We used fasting glucose summary statistics as a surrogate measure for the PRS of this measure.

**Lactose consumption computations.** We computed an estimate of average monthly lactose consumption (in grams), using a questionnaire of consumption frequency of 23 dairy products. As lactose consumption was exponentially distributed in our data, we log-transformed it to induce normality for the  $b^2$  and phenotype prediction analyses.

**Genetic kinship, principal components and relatedness estimation.** We used PC-Relate<sup>50</sup> for estimating genetic kinship and PC-AiR<sup>51</sup> for genetic principal components computation, as these tools are robust to the presence of relatedness and admixture. We used a filtered dataset of 75,384 SNPs in approximate linkage equilibrium ( $r^2 < 0.15$ ), and ran an iterative estimation procedure (with the initial kinship estimates provided by KING-Robust<sup>52</sup>) until the principal components computation converged, as previously described<sup>53</sup>. We estimated the degree of relatedness between individuals using their kinship coefficient and previously proposed cutoffs<sup>52</sup>. In the analysis of the LLD cohort and when testing kinship–ancestry associations, we used the kinship matrix estimated by GCTA<sup>24</sup>, as the kinship matrix of PC-Relate is by definition not associated with ancestry.

**Mantel tests.** Mantel tests used were performed with 100,000 permutations. When associating a matrix with a vector, we constructed a distance matrix for the vector using Euclidean distances. When not using covariates, we used a Spearman correlation-based Mantel test. In the presence of covariates, we performed a Pearson correlation-based partial Mantel test (performed by first regressing the matrix of covariate differences out of the two compared matrices, and then performing a Mantel test on the resulting residualized matrices), according to previous recommendations<sup>54</sup>. When testing association with specific taxa, we excluded taxa present in  $<5\%$  of individuals.

**Ancestry proportions prediction.** We attempted to predict ancestry proportions from microbiome composition using a variety of different techniques: Ridge regression<sup>55</sup>, lasso regression<sup>55</sup> and extreme gradient boosting<sup>56</sup>. We used as features either the top 100 PCOs of the Bray–Curtis dissimilarity matrix, the raw bacterial abundances (under various taxonomic levels) transformed to a logarithmic scale or the principal components of presence/absence of genes. Prediction accuracy was measured via a tenfold cross validation. The hyperparameters of the methods were determined in each fold via cross validation, using only the training set of each fold.

**Microbiome principal coordinates prediction.** We attempted to predict top microbiome PCOs from ancestry proportions or host genotypes via ridge regression, which is robust to high-dimensional data, with covariates used as additional explanatory variables. We computed  $P$  values via permutation testing with 10,000 permutations; in each permutation we assigned to each individual the microbiome PCOs of a random individual. The  $P$  value was defined as the number of permutations in which the sum of the coefficients of determination across the top 2, 5 or 10 PCOs was greater than that obtained under the non-permuted data.

**Analysis of data from twins study.** We estimated the overall microbiome heritability and the abundance of heritable taxa, using a previously published<sup>6</sup> dataset of 2,252 twins. Our analysis is based on two principles: First, we define the ‘overall microbiome heritability’ as a weighted average of taxa-specific heritabilities. The weight of each taxa is determined by its relative abundance in the TwinsUK data, and its heritability estimate is taken from the previous analysis<sup>6</sup>. Second, we assume that only a subset of bacterial taxa is heritable. Therefore, we include only a subset of taxa in the weighted average computation, corresponding to taxa with heritability  $P$  values (as previously computed<sup>6</sup>) smaller than a given cutoff.

We considered two cutoffs: a 5% FDR cutoff, and a liberal 5% false-positive rate cutoff, with no multiple testing corrections. The first cutoff probably yields a subset



of the truly heritable taxa, and the second cutoff probably yields a subset with all heritable taxa as well as many non-heritable taxa. The heritability estimates using these two subsets therefore serve as lower and upper bounds, respectively, on the overall microbiome heritability estimate.

Given a subset of taxa considered as significantly heritable, we estimated the overall heritability by using a weighted average of the estimated heritabilities of operational taxonomic units (OTUs) associated with these taxa, weighted by the relative abundances of these OTUs. The resulting quantity was then averaged across individuals. The estimated heritability of an OTU was the maximal heritability estimate among all heritable taxa with which it was associated.

**Testing for SNP–microbiome associations using the vegan package.** We repeated previously proposed<sup>9</sup> techniques for testing SNP–microbiome associations and estimating the variance inferred by several SNPs using the *envfit* and *ordiR2step* functions, respectively, in the *vegan* package in R<sup>54</sup>.

Permutation testing for the fraction of genus  $\beta$ -diversity variance that can be inferred from the top 42 SNPs (corresponding to the number used in the previous analysis<sup>9</sup>) was carried out as follows. We performed 10,000 permutation analyses. In each analysis we (i) randomly assigned to each individual the genotype of a randomly selected individual; (ii) ranked all SNPs according to their association with microbiome  $\beta$ -diversity, using the *envfit* function; and (iii) estimated the fraction of  $\beta$ -diversity variance that can be inferred from the combined top 42 SNPs, using the *ordiR2step* function. The resulting  $P$  value was the fraction of permutations in which the fraction of inferred variance was greater than observed under the real data.

**Testing for SNP associations with individual taxa.** We tested for associations between individual bacteria and individual SNPs using FaST-LMM<sup>38</sup>. We used all 814 genotyped individuals who passed quality control, including related individuals and individuals with a shared household, and controlled for these potential confounding sources using two variance components that encode kinship (as computed via PC-Relate<sup>50</sup>) and household sharing (using a binary co-sharing covariance matrix). When testing each SNP, we used the covariates described earlier, as well as the top five genetic principal components, and a genetic kinship matrix based only on SNPs from other chromosomes, to avoid proximal contamination<sup>57</sup>.

The abundance of bacteria present in at least 95% of individuals was encoded using the log-abundance; we excluded outlier individuals who were more than five standard deviations away from the mean. Otherwise, we dichotomized bacteria into presence/absence patterns and encoded the phenotype as a binary vector to prevent zero inflation, which leads to a bimodal distribution (LMMs handle binary phenotypes properly if the data are not ascertained<sup>58</sup>).

**Comparing results of different studies.** We evaluated the consistency of previous association studies<sup>5,6,8–10</sup> using the number of associations that are in the same locus (<100 kb apart) and associated with taxa belonging to the same phylum. We evaluated replication power by counting the number of SNPs in our own study with  $P < 0.05/211$  (corresponding to 211 previously reported loci), using the closest imputed SNP to the reported one.

**Relatives and household-sharing tests.** We tested for significant microbiome sharing among related individuals or individuals sharing a household, by comparing their average Bray–Curtis dissimilarity to that of pairs with no family relation or household sharing using a permutation test with 100,000 permutations. In each permutation, we randomly divided the combined set of all pairs into two disjoint sets while preserving the original set sizes, and asked whether the mean difference in Bray–Curtis dissimilarity between individuals in the two sets was greater than the difference observed in the real data. To prevent confounding effects, we considered only individuals whose stool was collected with a swab (one of the two stool collection methods).

**Associating environmental factors with the microbiome.** We tested for associations between 201 environmental factors and microbiome  $\beta$ -diversity at the species level with PERMANOVA<sup>21</sup>, using data from self-reported questionnaires<sup>12</sup> (Supplementary Table 17). To quantify the fraction of microbiome variance that could be inferred from environmental factors in combination we performed a greedy stepwise algorithm, in which at each iteration we added the environmental factor that contributed the greatest fraction of inferred variance to factors added in previous iterations. Before adding each factor, we permuted it 100 times and verified that its contribution was greater than in at least 55% of these permutations. If not, we stopped the algorithm. The statistical significance of the resulting estimate was evaluated using a permutation testing with 100,000 permutations, in which for each permutation we assigned all 201 environmental factors of each individual to a random individual, and then reran the entire analysis (including the feature selection procedure).

To perform the above procedure with SNPs instead of environmental factors, we first selected a set of SNPs in approximate linkage equilibrium that are maximally associated with the microbiome  $\beta$ -diversity, and then performed the analysis using these SNPs. Specifically, we first sorted the SNPs according to their fraction of

inferred  $\beta$ -diversity variance, and then iteratively selected 201 top-ranking SNPs (corresponding to the number of environmental variants) that are not within 200 kb of a previously selected SNP. We then reran the PERMANOVA analysis with the selected SNPs.

**Computing polygenic risk scores.** PRSs were computed using  $\hat{y}_i = \sum_{j \in R(c)} x_i^j \hat{b}^j$ , in which  $\hat{y}_i$  is the predicted phenotype,  $x_i^j$  is SNP  $j$  of individual  $i$ ,  $\hat{b}^j$  is the effect of SNP  $j$  reported in summary statistics and  $R(c)$  is the set of SNPs found in both the genotyping array and in the summary statistics with  $P < c$  for the cutoff  $c$ . The optimal  $c$  value was selected by searching over the grid [ $10^0$ ,  $3 \times 10^{-1}$ ,  $10^{-1}$ , ...,  $10^{-8}$ ] and finding the value maximizing the Spearman correlation between the true and predicted phenotypes. To prevent overfitting, the value of  $c$  used to compute the PRS of every individual was estimated using a subset of 90% of the data that did not include this individual. Similarly, when performing phenotype prediction, we estimated  $c$  using only individuals in the training set. SNPs were normalized to have a unit variance, according to their reported allele frequency. We used the original rather than the imputed set of SNPs, as we empirically verified that using the imputed set of SNPs in conjunction with linkage disequilibrium pruning did not improve prediction results. The list of summary statistics used is provided in Supplementary Table 20.

**Construction of a kinship matrix based on microbial genes.** We encoded the bacterial kinship of individuals  $i, j$  using  $\sum_k g_i^k g_j^k / n$ , in which  $k$  iterates over all genes present in >1% of individuals,  $g_i^k$  is the presence/absence indicators of gene  $k$  in individual  $i$  (using a relative abundance cutoff of  $10^{-6}$ , and normalized to have a zero mean and a unit variance), and  $n$  is the number of genes (1,360,337).

**Microbiome-association index estimation.** Microbiome-association index ( $b^2$ ) was estimated using GCTA<sup>24</sup>, a tool used in statistical genetics for estimating SNP-based genetic kinship. Instead of a matrix of host SNPs, as is commonly used in GCTA, we used a microbial genes-based kinship matrix. For all phenotypes (except lactose consumption), the covariates included the PRS of the investigated phenotype, the covariates described earlier and the top five genetic principal components. In the analysis of lactose consumption, we replaced the PRS with the SNPs rs4988235 and rs182549, which largely explain the genetic component of lactase persistence in European populations<sup>59</sup>.  $P$  values were computed using RL-SKAT<sup>60</sup> and confidence intervals were computed using FIESTA<sup>25</sup>. Outlier individuals with phenotypes more than five standard deviations away from the mean were excluded from the analysis.

We defined  $b^2$  estimation accuracy for a given covariance matrix using the average width of 95% confidence intervals (assuming that  $b^2$  is uniformly distributed in [0,1]). We estimated this quantity by invoking FIESTA 100 times with 100 different  $b^2$  values evenly spaced in the interval [0,1] and averaging the resulting 95% confidence interval widths.

**Analysis of data from the Wellcome Trust.** We computed confidence intervals for genetic heritability estimation using 5,652 previously described<sup>26</sup> control individuals from the Wellcome Trust National Blood Service and 1958 birth cohorts. SNPs with >0.5% missing data,  $P < 0.01$  for allele frequency difference between the two groups,  $P < 0.000005$  for deviation from Hardy–Weinberg equilibrium or minor allele frequency <1% were removed. The genetic kinship matrix was computed using GCTA<sup>24</sup> and confidence intervals were estimated using FIESTA<sup>25</sup>.

**Phenotype prediction.** Phenotype prediction was performed with an LMM<sup>61</sup>, using a kinship matrix based on presence/absence of genes, constructed as described in ‘Construction of a kinship matrix based on microbial genes’ (LMMs are mathematically equivalent to a ridge regression model that uses the principal components of the kinship matrix as covariates, and they reduce to linear regression when not using a kinship matrix). The covariates included age, sex, and daily median caloric, carbohydrate, fat and protein consumption. In some experiments we additionally included covariates for host genetic effects, represented either as PRS (for all phenotypes except lactose consumption) or as the SNPs rs4988235 and rs182549 for lactose consumption. Prediction performance was evaluated using a tenfold cross validation. Outlier individuals with phenotypes more than five standard deviations away from the mean were excluded from all analyses. We also evaluated additional types of kinship matrices: (i) a  $\beta$ -diversity matrix, which we transformed to a kinship matrix as previously described<sup>62</sup>; and (ii) kinship matrices based on relative abundances or presence/absence of bacterial taxa instead of genes (Supplementary Table 23).

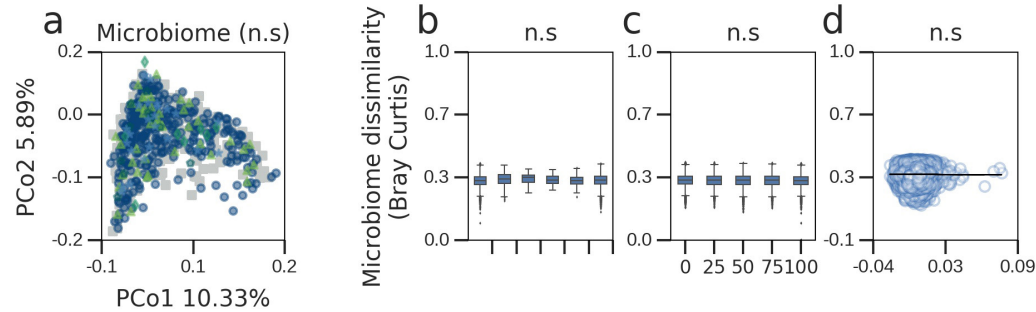
**Meta-analysis.** Meta-analysis of the Israeli and LLD cohorts was performed using Stouffer's method.

**Israeli cohort.** The Israeli cohort study was approved by Tel Aviv Sourasky Medical Center Institutional Review Board, approval numbers TLV-0658-12, TLV-0050-13 and TLV-0522-10; Kfar Shaul Hospital Institutional Review Board, approval number 0-73; and Weizmann Institute of Science Bioethics and Embryonic Stem Cell Research oversight committee. The study was reported to <http://clinicaltrials.gov/>, NCT number: NCT01892956.

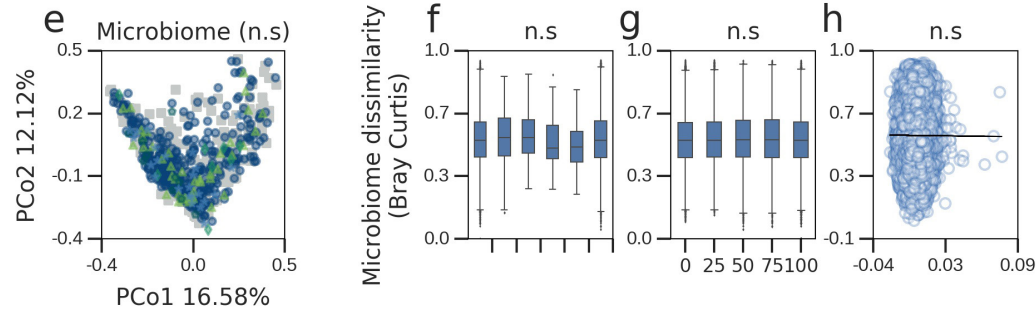
**Data availability.** The accession number for the datasets analysed in this paper are: (i) Israeli metagenome and 16S rRNA gene sequences, European Nucleotide Archive (ENA; <https://www.ebi.ac.uk/ena>): PRJEB11532; (ii) TwinsUK 16S rRNA gene sequences, ENA: ERP015317; (iii) LifeLines DEEP sequencing data, European Genome-phenome Archive (EGA; <https://www.ebi.ac.uk/ega/>): EGAS00001001704; and (iv) Wellcome Trust 2 genotypes, EGA: EGAD00000000021 and EGAD00000000023. All relevant data are available from the corresponding authors upon reasonable request. Source data for Fig. 3a and Extended Data Figs 1, 2 is available from [http://genie.weizmann.ac.il/genomica\\_links.html](http://genie.weizmann.ac.il/genomica_links.html).

39. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
41. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
42. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
43. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
44. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
45. Carmi, S. *et al.* Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat. Commun.* **5**, 4835 (2014).
46. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
47. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
48. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
49. Hong, C. *et al.* PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2**, 33 (2014).
50. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).
51. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* **39**, 276–293 (2015).
52. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
53. Conomos, M. P. *et al.* Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos. *Am. J. Hum. Genet.* **98**, 165–184 (2016).
54. Oksanen, J. *et al.* vegan: community ecology package. <https://cran.r-project.org/web/packages/vegan/index.html> (2017).
55. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2009).
56. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
57. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012).
58. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
59. Ingram, C. J. E., Mulcare, C. A., Itan, Y., Thomas, M. G. & Swallow, D. M. Lactose digestion and the evolutionary genetics of lactase persistence. *Hum. Genet.* **124**, 579–591 (2009).
60. Schweiger, R. *et al.* RL-SKAT: an exact and efficient score test for heritability and set tests. *Genetics* **207**, 1275–1283 (2017).
61. de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C. & Sorensen, D. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* **9**, e1003608 (2013).
62. Zhao, N. *et al.* Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* **96**, 797–807 (2015).

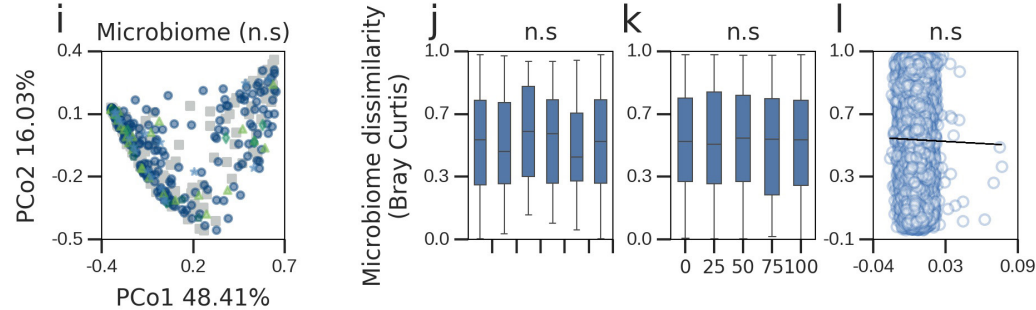
## Gene level



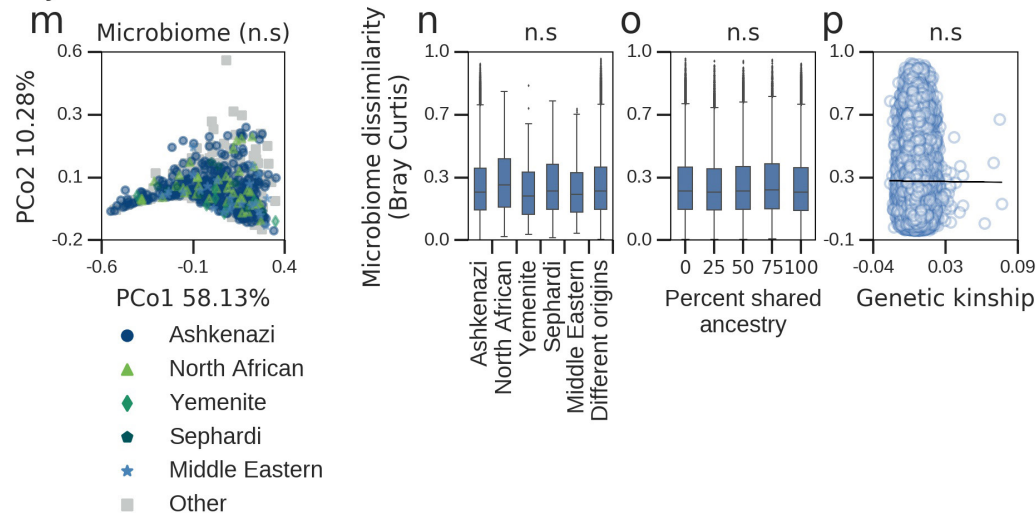
## Genus level



## 16S Genus level



## Phylum level

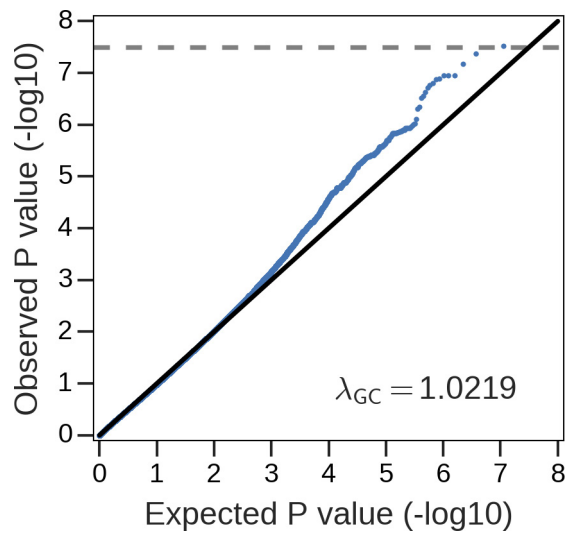


**Extended Data Figure 1 | Limited evidence for microbiome associations with genetic ancestry or kinship across multiple functional and taxonomic levels.** **a–p.** Each row is similar to Figs 1b, d–e, 2b, but is based on the abundance of bacterial genes (**a–d**), genera (**e–h**), genera based on 16S rRNA gene sequencing data (**i–l**) or phyla (**m–p**).

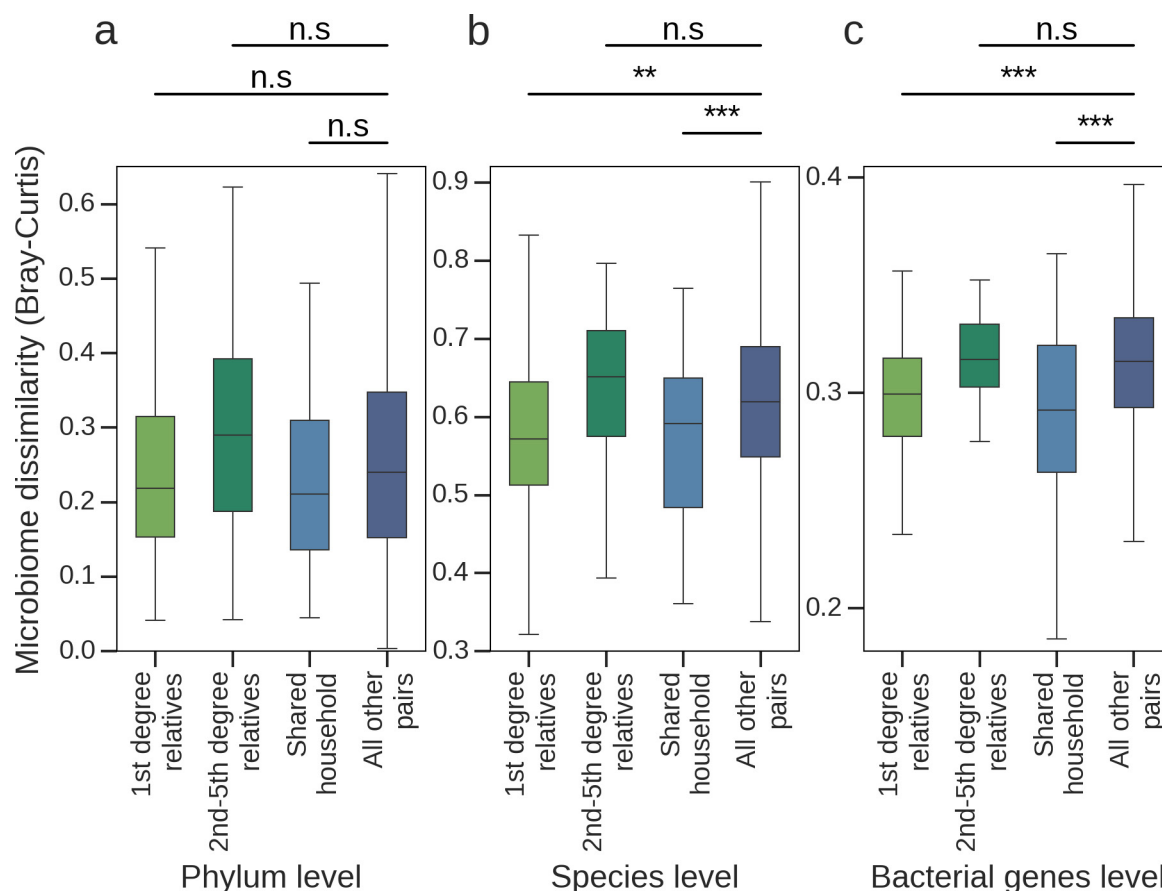
**a, d, e, h, m, p,**  $n = 715$  genotyped individuals; **i, l,**  $n = 481$  individuals

with 16S rRNA gene sequencing data; **b, f, n,**  $n = 737$  individuals for whom the ancestries of all grandparents are known; **j,**  $n = 509$  individuals with 16S rRNA gene sequencing data for whom the ancestries of all grandparents are known; **c, g, o,**  $n = 946$  individuals; and **k,**  $n = 650$  individuals with 16S rRNA gene sequencing data.



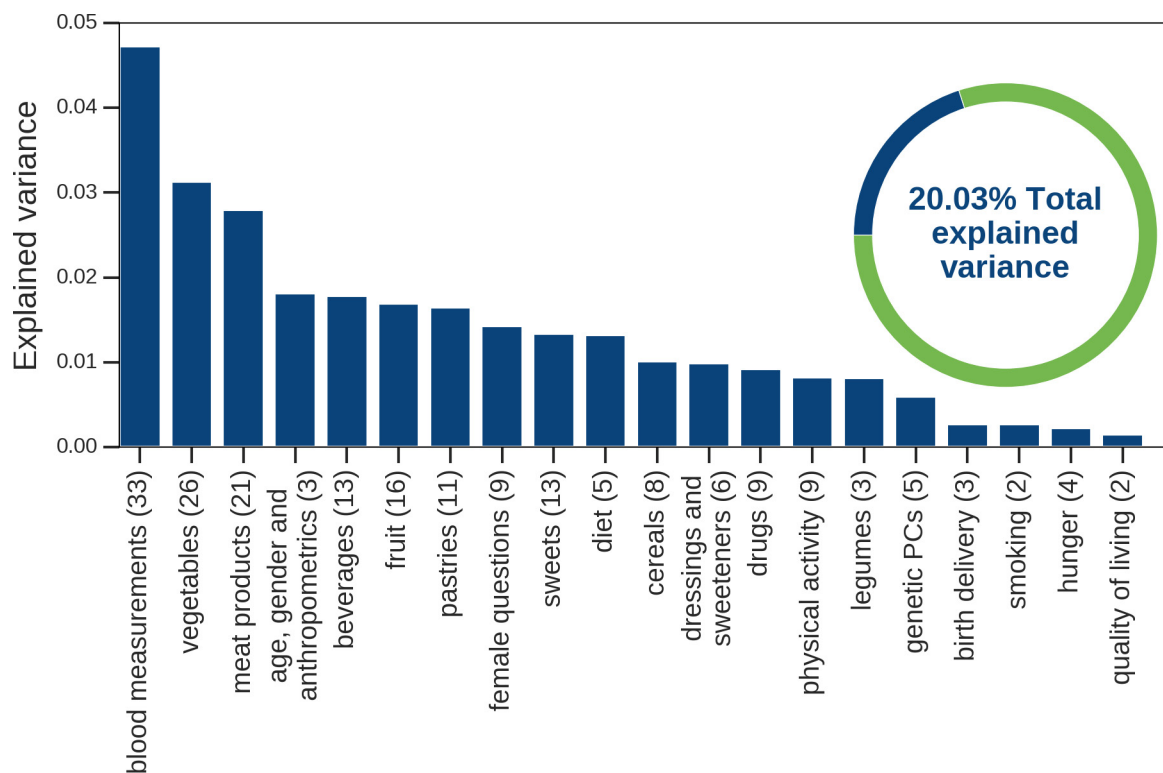


**Extended Data Figure 2 | Limited evidence for associations between microbiome  $\beta$ -diversity and specific SNPs.** The quantile–quantile plot shows that only two SNPs are significantly associated with microbiome  $\beta$ -diversity at  $P < 5 \times 10^{-8}$ , computed using a distance-based  $F$  test with  $n = 715$  unrelated genotyped individuals.  $\lambda_{GC}$ , genomic inflation factor.



**Extended Data Figure 3 | Individuals who share a household at present or have shared one in the past have significantly similar microbiomes.** First-degree relatives and individuals with present household sharing have significantly similar species and bacterial gene abundances ( $P < 0.01$ ; permutation testing). **a–c**, Box plots depict the distribution of Bray–Curtis dissimilarities across pairs of individuals at the phylum (**a**), species (**b**) and bacterial genes (**c**) level. Each panel shows the Bray–Curtis dissimilarities among all pairs of (i) first-degree relatives, who are likely to

have experienced present or past household sharing ( $n = 55$  pairs); (ii) second-to-fifth-degree relatives, who are unlikely to have experienced present or past household sharing ( $n = 24$  pairs); (iii) unrelated individuals self-reported to currently share a household ( $n = 32$  pairs); and (iv) all other individuals ( $n = 255,891$  pairs). The lower and upper limits of the boxes represent the 25% and 75% percentiles, respectively, and the top and bottom whiskers represent the 5% and 95% percentiles, respectively. The  $P$  value ranges for all panels are:  $**P < 0.01$  and  $***P < 0.005$ .

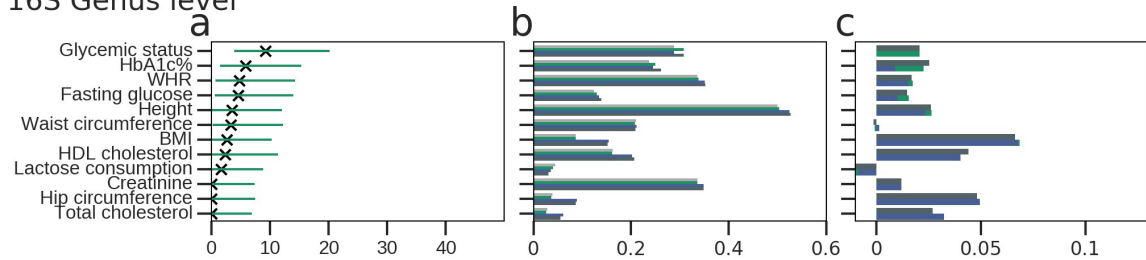


**Extended Data Figure 4 | The gut microbiome is significantly associated with multiple environmental factors.** The fraction of variance of the microbiome  $\beta$ -diversity matrix that can be inferred from different categories of environmental factors is shown.  $n = 715$  individuals (Supplementary Table 17); numbers in parentheses indicate the number of

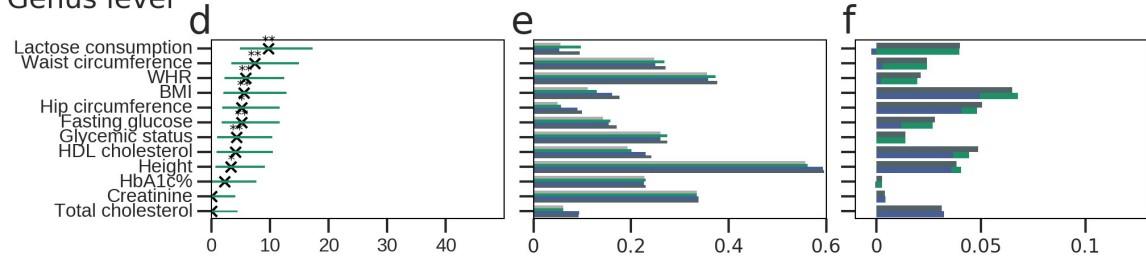
features in each category. The fraction of inferred variance can reflect both the information that the category conveys on the microbiome as well as the number of factors in the category, which depends on the questionnaire used in the study.



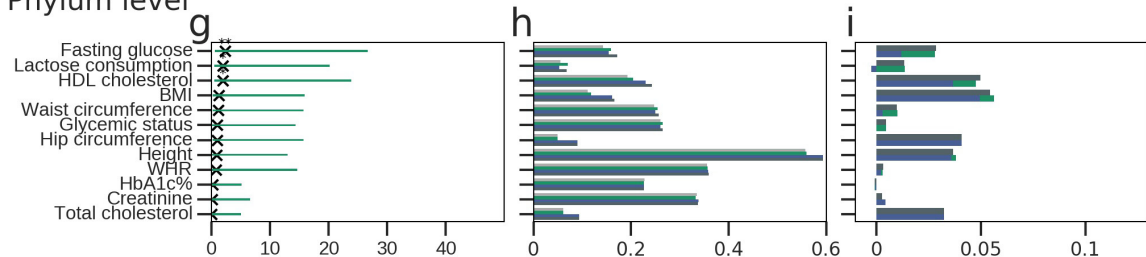
## 16S Genus level



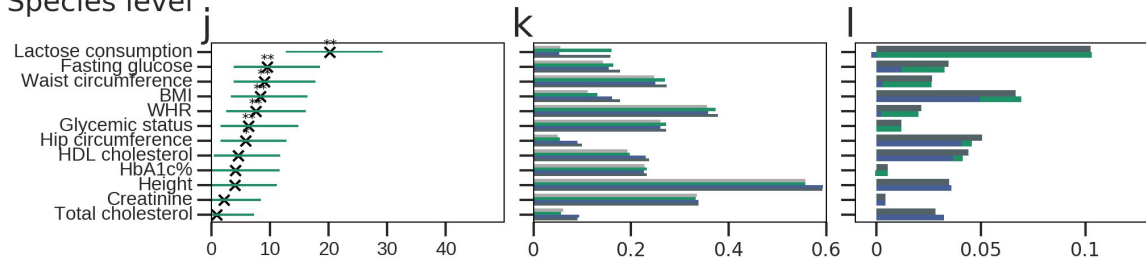
## Genus level



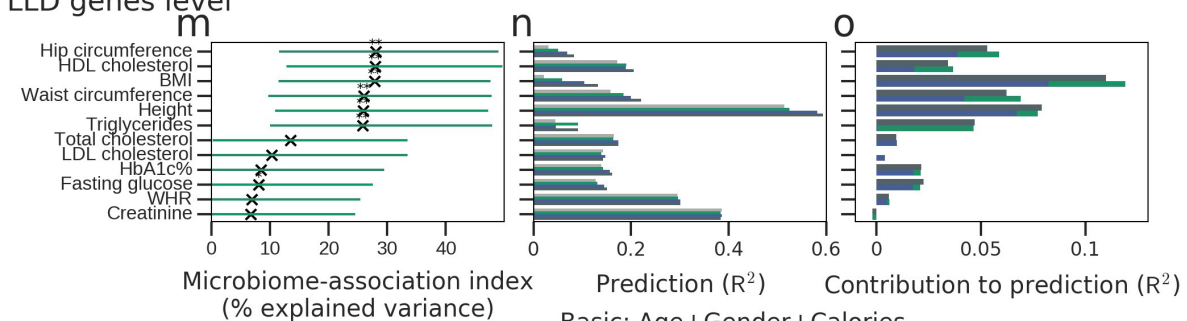
## Phylum level



## Species level



## LLD genes level

Microbiome-association index  
(% explained variance)Prediction ( $R^2$ )Contribution to prediction ( $R^2$ )

Basic: Age+Gender+Calories

Basic

Basic+Microbiome

Basic+Genetics

Basic+Genetics

Basic+Genetics

+Microbiome

Microbiome

Genetics

Genetics+Microbiome

**Extended Data Figure 5 |  $b^2$  estimates and phenotype prediction results when using various data sources.** Each row is similar to Fig. 4c–e, but is based on a different data source. **a–c**, Relative abundance of genera, obtained from 16S rRNA gene sequencing (using  $n = 464$  individuals). **d–f**, Relative abundance of genera, obtained from metagenomic sequencing (using  $n = 715$  individuals). **g–i**, Relative abundance of phyla (using  $n = 715$  individuals). **j–l**, Relative abundance of species (using

$n = 715$  individuals). **m–o**, Relative abundance of bacterial genes in the LLD cohort (using  $n = 836$  individuals). Note that two phenotypes that were analysed in the Israeli cohort (lactose consumption and glycaemic status) were not available for the LLD cohort, and two phenotypes available for the LLD cohort and shown here (LDL cholesterol and triglycerides) were not available for the Israeli cohort. The  $P$  value ranges for all panels are: \*FDR < 0.05, \*\*FDR < 0.01 and \*\*\*FDR < 0.001.

**Extended Data Table 1 | Baseline characteristics of the cohort**

Property	Value (mean±SD)
Number of participants	1,046
Age	42.6±12.7
Gender (%female)	61%
Median calories consumed daily (kcal)	1867±530
Median carbohydrates consumed daily (gr)	210.7±66.1
Median fat consumed daily (gr)	74.2±27.6
Median protein consumed daily (gr)	69.1±24.8
BMI	26.6±5.2
Waist circumference (cm)	87.7±13.6
Hip Circumference (cm)	104.6±13.6
Waist-hip ratio	0.84±0.09
Height (cm)	167±9.1
Total cholesterol (mg/dl)	186±37
HDL cholesterol (mg/dl)	58.1±17.1
HbA1c%	5.44±0.47
Fasting glucose (mg/dl)	92.05±11.3
Creatinine (mg/dl)	0.84±0.19
Lactose consumption (gr)	562±604

The mean and standard deviation of all properties used as covariates or as investigated phenotypes are shown. Dietary properties are based on information recorded in real time by study participants on their smartphones (see Methods).

**Extended Data Table 2 | No significant association between ancestral or genetic similarity and the gut microbiome**

		Ancestry (categorical)	Ancestry (proportions)	Genetic kinship
Host genetics (control)	Kinship	<0.0001	<10 <sup>-5</sup>	-
	$\beta$ -diversity	>0.06	0.73	0.59
Microbiome	$\alpha$ -diversity	0.57	0.22	0.61
	Specific taxa	>0.05	>0.05	>0.05

Each cell contains the *P* value of a single or of multiple statistical tests, testing whether individuals who are more similar according to ancestry or genetic kinship (in columns) are also more similar according to (i) microbiome  $\beta$ -diversity (using Bray–Curtis dissimilarity); (ii) microbiome  $\alpha$ -diversity (using Shannon diversity); (iii) abundance of specific taxa; or (iv) genetic kinship (in rows). The first column includes *n* = 582 non-admixed individuals, the second includes 946 individuals and the third includes 715 unrelated genotyped individuals. *P* values in the first column are based on Kruskal–Wallis tests (using the top 5 microbiome PCOs for Bray–Curtis dissimilarity, and the top 5 genetic principal components for genetic kinship); *P* values in the other columns are based on Mantel tests (Methods).



# Clusters of cyclones encircling Jupiter's poles

A. Adriani<sup>1</sup>, A. Mura<sup>1</sup>, G. Orton<sup>2</sup>, C. Hansen<sup>3</sup>, F. Altieri<sup>1</sup>, M. L. Moriconi<sup>4</sup>, J. Rogers<sup>5</sup>, G. Eichstädt<sup>6</sup>, T. Momary<sup>2</sup>, A. P. Ingersoll<sup>7</sup>, G. Filacchione<sup>1</sup>, G. Sindoni<sup>1</sup>, F. Tabataba-Vakili<sup>2</sup>, B. M. Dinelli<sup>4</sup>, F. Fabiano<sup>4,8</sup>, S. J. Bolton<sup>9</sup>, J. E. P. Connerney<sup>10</sup>, S. K. Atreya<sup>11</sup>, J. I. Lunine<sup>12</sup>, F. Tosi<sup>1</sup>, A. Migliorini<sup>1</sup>, D. Grassi<sup>1</sup>, G. Piccioni<sup>1</sup>, R. Noschese<sup>1</sup>, A. Cicchetti<sup>1</sup>, C. Plainaki<sup>13</sup>, A. Olivieri<sup>13</sup>, M. E. O'Neill<sup>14</sup>, D. Turrini<sup>1,15</sup>, S. Stefani<sup>1</sup>, R. Sordini<sup>1</sup> & M. Amoroso<sup>13</sup>

**The familiar axisymmetric zones and belts that characterize Jupiter's weather system at lower latitudes give way to pervasive cyclonic activity at higher latitudes<sup>1</sup>. Two-dimensional turbulence in combination with the Coriolis  $\beta$ -effect (that is, the large meridionally varying Coriolis force on the giant planets of the Solar System) produces alternating zonal flows<sup>2</sup>. The zonal flows weaken with rising latitude so that a transition between equatorial jets and polar turbulence on Jupiter can occur<sup>3,4</sup>. Simulations with shallow-water models of giant planets support this transition by producing both alternating flows near the equator and circumpolar cyclones near the poles<sup>5–9</sup>. Jovian polar regions are not visible from Earth owing to Jupiter's low axial tilt, and were poorly characterized by previous missions because the trajectories of these missions did not venture far from Jupiter's equatorial plane. Here we report that visible and infrared images obtained from above each pole by the Juno spacecraft during its first five orbits reveal persistent polygonal patterns of large cyclones. In the north, eight circumpolar cyclones are observed about a single polar cyclone; in the south, one polar cyclone is encircled by five circumpolar cyclones. Cyclonic circulation is established via time-lapse imagery obtained over intervals ranging from 20 minutes to 4 hours. Although migration of cyclones towards the pole might be expected as a consequence of the Coriolis  $\beta$ -effect, by which cyclonic vortices naturally drift towards the rotational pole, the configuration of the cyclones is without precedent on other planets (including Saturn's polar hexagonal features). The manner in which the cyclones persist without merging and the process by which they evolve to their current configuration are unknown.**

NASA's Juno spacecraft<sup>10,11</sup> has been operating in a 53-day highly elliptical polar orbit of Jupiter since 5 July 2016. The spacecraft has passed close to Jupiter six times now, on five of which occasions instruments on board were able to sound the planet and observe many interesting atmospheric structures<sup>12–15</sup>. The Juno spacecraft is in a high-inclination orbit with perijove (the point in its orbit nearest Jupiter's centre) approximately 4,000 km above the cloud tops, passing from pole to equator to pole in about two hours. From their unique vantage point above the poles, JIRAM<sup>16,17</sup> (Jupiter InfraRed Auroral Mapper) and JunoCam<sup>18</sup>, onboard Juno, obtained unprecedented views of Jupiter's polar regions. JIRAM is an infrared imager suitable for atmospheric mapping and JunoCam is a pushframe visible camera. Jupiter fly-bys took place during perijove passes PJ1 on 28 August 2016, PJ3 on 11 December 2016, PJ4 on 2 February 2017 and PJ5 on 27 March 2017 (no remote-sensing observations were collected during PJ2).

The atmospheric structure in Jupiter's polar regions is very different from the well known axisymmetric banding of alternating belts and

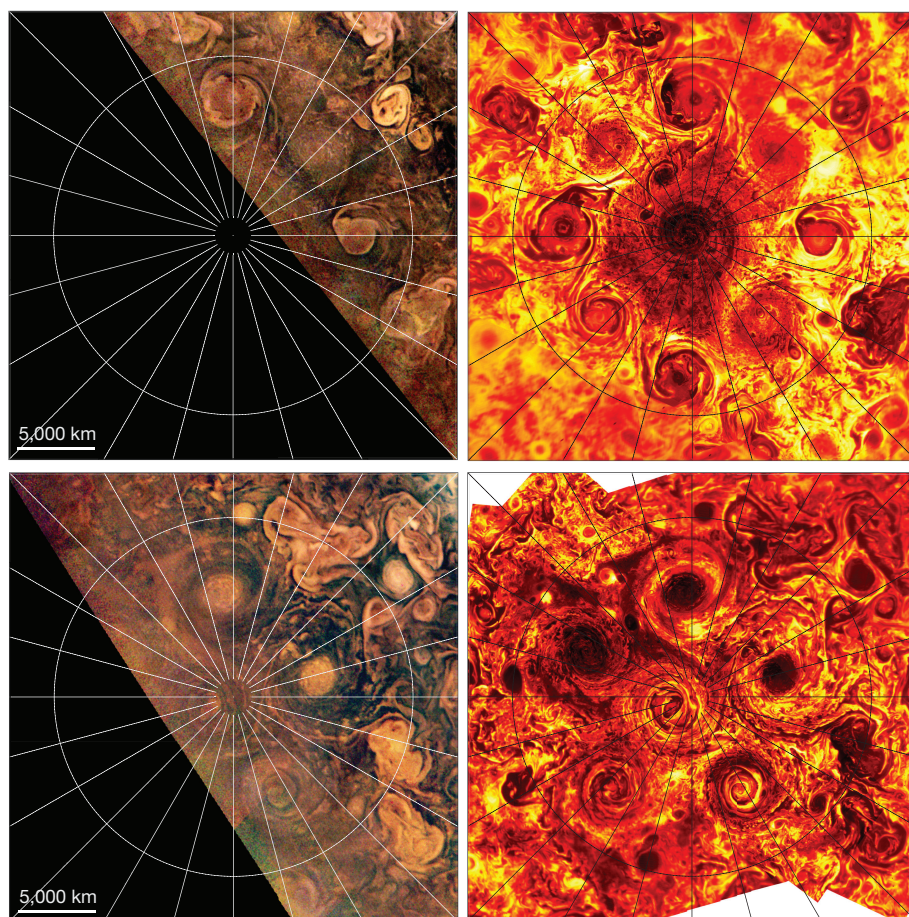
zones at lower latitudes. The polar turbulence predicted by models is consistent with initial close-up observations in the visible part of the spectrum<sup>15</sup>. Cyclones, as opposed to anticyclones, were expected in the polar regions as a result of the Coriolis  $\beta$ -effect<sup>9,19,20</sup>. What was unexpected is their stable appearance, close clustering and symmetry around each of the poles.

The Northern Polar Cyclone (NPC, Fig. 1) has a diameter of approximately 4,000 km (on the JIRAM infrared images). It is offset relative to the geographic north pole of Jupiter by about 0.5° and is surrounded by eight circumpolar cyclones (henceforth referred to as just 'cyclones') in a double-squared geometric pattern (Figs 1 and 2). Counting alternating cyclones, four are centred at about 83.3° N and the other four are centred at about 82.5° N. The square formed by the latter four cyclones is shifted with respect to the square formed by the former four cyclones by 45° longitude, forming a 'ditetragonal' shape, in which the angular distances between the centre of one cyclone and the next vary from 43° to 47°. All cyclones have similar dimensions with diameters ranging from 4,000 km to 4,600 km. Spiral arms are prominent in their outer regions, but tend to disappear in their inner regions except in the NPC itself. These arms define an additional sphere of influence beyond the cores of the cyclones in which co-rotating material can be found. The four cyclones furthest from the NPC have broad cloud-covered inner regions with sharp oblate boundaries. The four cyclones interspersed between them have more diverse and irregular inner regions, with very small-scale cloud textures; some of them appear chaotic and turbulent.

The Southern Polar Cyclone (SPC, Figs 1 and 2) is surrounded by five large circumpolar cyclones in a quasi-pentagonal pattern. They are of similar size, but are generally bigger than the northern cyclones, with diameters ranging between 5,600 km and 7,000 km. The southern cyclones present a range of morphologies, although the differences are much less distinct than in the north. In particular, some of them display a quasi-laminar circulation: the SPC and two adjacent cyclones have cloud spirals converging to the centre, while the other three cyclones appear to be very turbulent along their spiral cloud branches. The SPC has an offset of about 1°–2° relative to the geographic south pole and the angular distance between two adjacent cyclones is not as regular as in the north: it can vary from 65° to 80° relative to the centre of rotation of the SPC.

Figures 1 and 2 show the correspondence between the features in JIRAM maps and in JunoCam images. Regions that are relatively bright in the JunoCam images are cool in the JIRAM thermal infrared images and regions that are relatively dark in the visible are warm. Because the JIRAM thermal radiance in the approximately 5- $\mu$ m M-band is primarily governed by cloud opacity, regions that appear warm can be interpreted as relatively clear of clouds, allowing radiance from deeper, warmer regions to be detected, and regions that appear cold must be cloudier.

<sup>1</sup>INAF-Istituto di Astrofisica e Planetologia Spaziali, Roma, Italy. <sup>2</sup>Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA. <sup>3</sup>Planetary Science Institute, Tucson, Arizona, USA. <sup>4</sup>CNR-Istituto di Scienze dell'Atmosfera e del Clima, Bologna e Roma, Italy. <sup>5</sup>British Astronomical Association, Burlington House, Piccadilly, London W1J 0DU, UK. <sup>6</sup>Alexanderstraße 21, 70184 Stuttgart, Germany. <sup>7</sup>Division of Geology and Planetary Sciences, California Institute of Technology, Pasadena, California, USA. <sup>8</sup>Dipartimento di Fisica e Astronomia, Università di Bologna, Bologna, Italy. <sup>9</sup>Space Science and Engineering Division, Southwest Research Institute, San Antonio, Texas, USA. <sup>10</sup>Code 695, NASA/Goddard Space Flight Center, Greenbelt, Maryland, USA. <sup>11</sup>Planetary Sciences Laboratory, University of Michigan, Ann Arbor, Michigan, USA. <sup>12</sup>Center for Astrophysics and Space Science, Cornell University, Ithaca, New York, USA. <sup>13</sup>Agenzia Spaziale Italiana, Roma, Italy. <sup>14</sup>Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois, USA. <sup>15</sup>Departamento de Física, Universidad de Atacama, Copayapu 485, Copiapó, Chile.

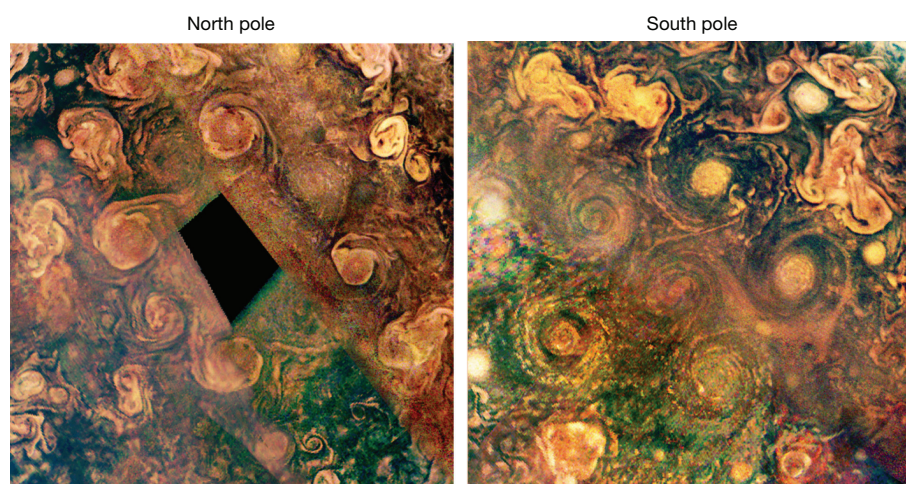


**Figure 1 | The poles of Jupiter as they appear at visible and infrared wavelengths.** Projected maps of the regions surrounding the north pole (top) and south pole (bottom) from the JIRAM 5- $\mu\text{m}$  M-filter observations (right panels) and JunoCam colour-composite images (left panels) during PJ4 on 2 February 2017. The latitude circle is 80° N or 80° S (planetocentric). Meridians are drawn every 15° of longitude, and 0° W in System III is positioned at the centre right of the images. By operating at thermal-infrared wavelengths, JIRAM observes the atmospheric structures regardless of solar illumination, whereas JunoCam's optical images are restricted to only the illuminated hemisphere, which is why only part of the JunoCam map for the north pole is present. JIRAM radiance, ranging from 0.02  $\text{W m}^{-2} \text{sr}^{-1}$  (dark red) to 0.8  $\text{W m}^{-2} \text{sr}^{-1}$  (white) is corrected with respect to the emission angle; the radiance scale is logarithmic. The JunoCam images are corrected with respect to solar illumination angle, as discussed in ref. 5 and the colours of the maps have been stretched and balanced to enhance atmospheric features. Cyclonic features can be seen clustered around each pole with regular circular shapes, some with spiral arms. For the south polar region, we note that there is a wider longitude separation (a 'gap') between the cyclones near 180° W (centre left side) than between the other cyclones. Two smaller cold (dark red) features can be seen to the upper left of the NPC, which are anticyclonic vortices.

Thus, the visibly bright discrete features in the JunoCam images in Figs 1 and 2 correspond to high-altitude clouds, while the darker background corresponds to a deeper cloud deck. This corresponds to a general qualitative result from JunoCam observations made during PJ1, that visually bright regions correspond to regions that are also relatively bright in the 890-nm band, which is sensitive to absorption by methane gas, implying

high-altitude clouds in those regions<sup>14</sup>. Figure 3, with the highest-resolution maps of the polar regions, gives a detailed view of the polar morphologies, showing JIRAM images corresponding to brightness temperatures in the range 190–260 K.

In most cases, the cyclones are essentially in contact if the spiral arms that extend beyond the core are included. In some cases, a single



**Figure 2 | The poles observed by JunoCam during the first four passes at Jupiter.** A composite is shown of the polar regions observed by JunoCam not only during PJ4 but also at complementary longitudes, acquired during PJ1, PJ3 and PJ5 for regions not illuminated by sunlight during PJ4. The PJ4 projection has been preserved as in the left panels in Fig. 1; the remainder of the unfilled space is covered by a composite of images from the other perijove passes. The remaining regions that are dark in the left panels of Fig. 1 are a smooth composite of JunoCam images taken during PJ1, PJ3 and PJ5. The area in the centre of the north polar

region (left panel) is dark because those latitudes were not illuminated. Elsewhere on Jupiter, cyclonic circulations assume various forms, especially at high latitudes, but none is a simple spiral with a circular outline, except for some very small ones. We note that, although they were imaged 53–106 days (1–2 Juno orbits) from the PJ4 observations, the positions and even the gross morphologies of the cyclones imaged during those orbits are not very different from their overall morphology in the PJ4 JIRAM map. The JunoCam map colours were chosen to enhance atmospheric features.



**Table 1 | Spin velocities of the single cyclones**

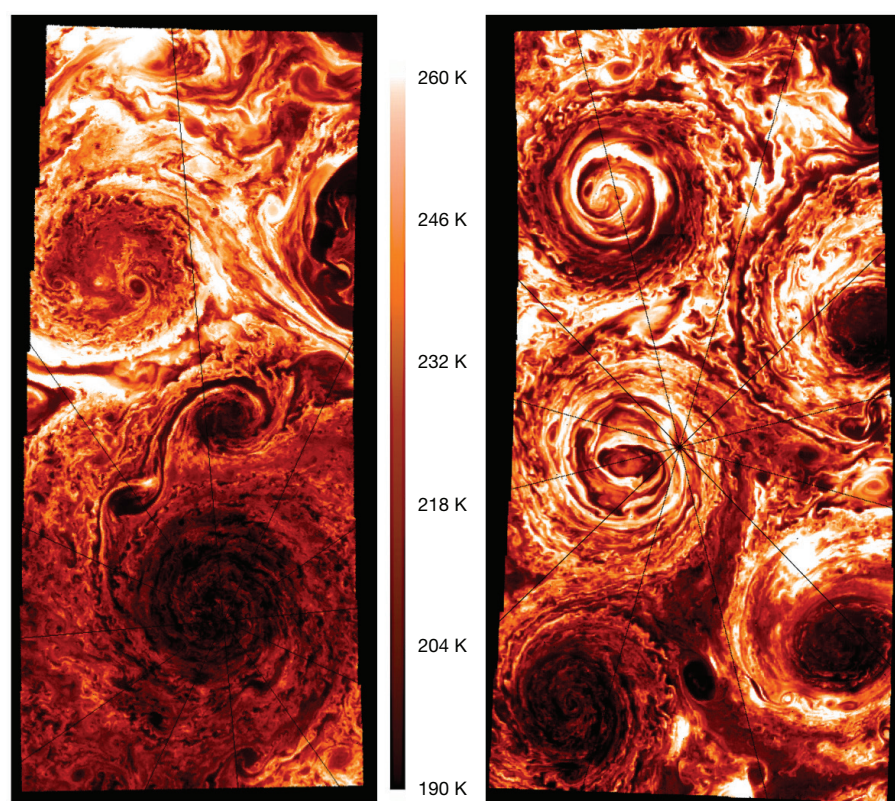
North pole				South pole			
Circumpolar cyclone number	Angular velocity (deg min <sup>-1</sup> )	Tangent velocity (km h <sup>-1</sup> )	Full rotation hours	Circumpolar cyclone number	Angular velocity (deg min <sup>-1</sup> )	Tangent velocity (km h <sup>-1</sup> )	Full rotation hours
NPC	0.17	267	35.3	SPC	0.19	299	31.6
1	0.22	343	27.5	1 (H)	0.13	204	46.1
3 (D)	0.21	337	28.0	2	0.17	273	34.5
4 (E)	0.10	157	60.0	3	0.16	252	37.5
5 (F)	0.12	296	48.0	4	0.21	330	28.6
6	0.19	295	32.0	5 (G)	0.17	273	34.5
7	0.22	354	26.6				
8	0.12	296	48.0				

Spin velocities were calculated at the radial distance of approximately 1,500 km from the spinning centres during PJ4 by JIRAM images. There were not enough data to compute the velocity of northern circumpolar cyclone 2. The numbering of the northern circumpolar cyclones goes from 1 to 8 proceeding anticlockwise, with circumpolar cyclone 1 the one located at 0° longitude. The numbering of the southern circumpolar cyclones starts from the circumpolar cyclone at 150° longitude and proceeds anticlockwise. Letters in parentheses identify the circumpolar cyclone, if present, in Extended Data Fig. 1. A quick calculation assuming the gradient wind balance, which includes Coriolis, centrifugal and pressure forces, indicates pressure gradients of about 5–10 Pa km<sup>-1</sup> at 1,500 km from the circumpolar cyclone centres.

cloud streak connects the outer spiral arms of adjacent cyclones and can be seen to be continuously stretched. The SPC is in contact with the five cyclones around it. By contrast, the PJ4 JIRAM animation (see Supplementary Videos) reveals a chaotic zone between the NPC and the eight surrounding cyclones, within which there is a largely continuous westward (clockwise) flow at about 86° N; poleward of this, the eastward (anticlockwise) flow of the NPC begins. This chaotic zone appears to contain turbulent small-scale cloud textures and a few small anticyclonic vortices. The largest of these, located between the NPC and the cyclones, may be identical to a similar anticyclonic vortex at PJ5, having moved westward by 31° longitude in Juno's 53-day orbital period. JIRAM data acquired during PJ4 cover a time span of about 2 h at each pole, enabling us to monitor the movements of the clouds and other structures that are evident within each cyclone, which in turn permits the identification of cyclonic and anticyclonic zones. The velocity field inside each cyclone is not straightforward to evaluate from these data, both because the pointing inaccuracy of JIRAM is not negligible when dealing with fine-scale structures inside cyclones, and because detailed structures whose movement is visible are not scattered uniformly. Table 1 provides a summary of preliminary

JIRAM measurements of the rotational speeds of individual cyclones at 1,500 km from their respective centres.

The changes occurring in the polar polygons can be seen by the JIRAM observations with a time lapse of about 53 terrestrial days between PJ4 and PJ5 (Extended Data Fig. 1). By this analysis, the northern eight cyclones appear to drift very slowly around the north pole (or around the NPC), by approximately 2.6° eastward in System III longitude. The changes between the observations taken by JunoCam are based on images of the sunlit side of the poles during PJ1, PJ3, PJ4 and PJ5. On the other hand, JunoCam polar images have an overlap of about 90°–180° in sunlit longitudes between successive perijove observations in similar 53-day intervals, and can also be compared with the complete map from JIRAM at PJ4. They show that the eight cyclones are preserved throughout the entire seven-month period, retaining their individual morphological characteristics, and showing only minor movements (Fig. 2). The visible sector of the octagon rotated around the north pole as follows (positive is westward): PJ1 to PJ3, about +2°; PJ3 to PJ4, -4° to -7.5°; PJ4 to PJ5, 0° to -3.5°. Thus the octagon has not shown any progressive rotation about the pole in System III longitudes. Both instruments observe small meridional displacements



**Figure 3 | High-resolution view of the polar vortices.** The left panel shows the north pole as seen in the 5- $\mu$ m spectral region (JIRAM M-filter) at an average spatial resolution of 18 km per pixel. The right panel shows the south pole at an average spatial resolution of 25 km per pixel in the same filter. These maps represent the highest available spatial resolution of JIRAM images during PJ4. The red colour scale from black to white is associated with the apparent brightness temperature shown, covering 190–260 K. Some cyclones look more clearly structured with alternating cold (cloudier) and warm (clearer) banding as a function of radius. It also clearly depicts the mesoscale dynamics over Jupiter's polar regions, showing a chaotic environment with many wavy structures and smaller anticyclones and cyclones developing among the largest ones. Such small anticyclonic eddies can be seen between some of the cyclones, especially around the NPC, where the largest of them measures about 1,200 km in diameter. There is a great structural difference between the NPC, which is dominated by a very small-scale cloud structure, and the SPC, which is characterized by a quasi-laminar behaviour. The SPC has a diameter of about 5,800 km and its centre is very peculiar, presenting an elongated 'eye' shape instead of the circular structure characterizing the centre of all of the other cyclones.



of individual cyclones of the same order of magnitude. For the cyclones at the south pole, JunoCam comparisons between perijove passes do suggest a progressive anticlockwise zonal rotation relative to the SPC of  $+1^\circ$  every 53 days, as well as some wandering of individual cyclones. There are large variations in the spacing of the cyclones around the pentagon, associated with the opening and closing of a gap that is always present between two of the cyclones. Just as in the north, the cyclones have preserved their individual morphologies over the seven months of observations.

Two questions arise from these data. The first is why the pentagon and octagon drift so slowly or not at all. By Stokes' theorem, net cyclonic vorticity at the centre would imply cyclonic circulation around the periphery. The other question is why the vortices do not merge. Saturn has a single cyclonic vortex at each pole. By analysing the conditions for formation of each Saturn vortex and comparing them with the conditions on Jupiter, it was predicted that the polar circulation could be different on Jupiter<sup>9</sup>. Some studies<sup>21</sup>, applying the theory to the merger of Jupiter's white ovals in 1998–2000, have also shown that like-signed vortices merge on a fast, advective timescale of four months when they are no longer separated by opposite-signed vortices in a single path, a 'vortex street'. Mergers of the polar cyclones are possible, but they have not occurred over seven months of observation, nor is there any evidence of new structures appearing inside the cyclone polygons. Finally, on the other hand, other studies<sup>22,23</sup> show that polygonal vortex patterns (vortex crystals) can develop owing to interaction with a background of weaker vorticity and last indefinitely in a two-dimensional Euler flow.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 24 July; accepted 15 November 2017.**

- Porco, C. C. *et al.* Cassini imaging of Jupiter's atmosphere, satellites, and rings. *Science* **299**, 1541–1547 (2003).
- Rhines, P. B. Waves and turbulence on a beta-plane. *J. Fluid Mech.* **69**, 417–443 (1975).
- Theiss, J. Equatorward energy cascade, critical latitude, and the predominance of cyclonic vortices in geostrophic turbulence. *J. Phys. Oceanogr.* **34**, 1663–1678 (2004).
- Sayanagi, K. M., Showman, A. P. & Dowling, T. E. The emergence of multiple robust zonal jets from freely evolving, three-dimensional stratified geostrophic turbulence with applications to Jupiter. *J. Atmos. Sci.* **65**, 3947–3962 (2008).
- Cho, J. Y. K. & Polvani, L. M. The emergence of jets and vortices in freely evolving, shallow-water turbulence on a sphere. *Phys. Fluids* **8**, 1531–1552 (1996).
- Iacono, R., Struglia, M. V. & Ronchi, C. Spontaneous formation of equatorial jets in freely decaying shallow water turbulence. *Phys. Fluids* **11**, 1272–1274 (1999).
- Showman, A. P. Numerical simulations of forced shallow-water turbulence: effects of moist convection on the large-scale circulation of Jupiter and Saturn. *J. Atmos. Sci.* **64**, 3132–3157 (2007).
- Scott, R. K. & Polvani, L. M. Forced-dissipative shallow-water turbulence on the sphere and the atmospheric circulation of the giant planets. *J. Atmos. Sci.* **64**, 3158–3176 (2007).
- O'Neill, M. E., Emanuel, K. A. & Flierl, G. R. Polar vortex formation in giant-planet atmospheres due to moist convection. *Nat. Geosci.* **8**, 523–526 (2015).
- Bolton, S. J. *et al.* Jupiter's interior and deep atmosphere: the initial pole-to-pole passes with the Juno spacecraft. *Science* **356**, 821–825 (2017).
- Connerney, J. E. P. *et al.* Jupiter's magnetosphere and aurorae observed by the Juno spacecraft during its first polar orbits. *Science* **356**, 826–832 (2017).
- Grassi, D. *et al.* Preliminary results on the composition of Jupiter's troposphere in hot spot regions from the JIRAM/Juno instrument. *Geophys. Res. Lett.* **44**, 4615–4624 (2017).
- Sindoni, G. *et al.* Characterization of the white ovals on Jupiter's southern hemisphere using the first data by the Juno/JIRAM instrument. *Geophys. Res. Lett.* **44**, 4660–4668 (2017).
- Orton, G. S. *et al.* The first close-up images of Jupiter's polar regions: results from the Juno mission JunoCam instrument. *Geophys. Res. Lett.* **44**, 4599–4606 (2017).
- Orton, G. S. *et al.* Multiple-wavelength sensing of Jupiter during the Juno mission's first perijove passage. *Geophys. Res. Lett.* **44**, 4607–4614 (2017).
- Adriani, A. *et al.* JIRAM, the Jovian Infrared Auroral Mapper. *Space Sci. Rev.* **213**, 393–446 (2017).
- Adriani, A. *et al.* Juno's Earth flyby: the Jovian Infrared Auroral Mapper preliminary results. *Astrophys. Space Sci.* **361**, 272 (2016).
- Hansen, C. J. *et al.* JunoCam: Juno's Outreach Camera. *Space Sci. Rev.* **213**, 475–506 (2017).
- Theiss, J. A generalized Rhines effect and storms on Jupiter. *Geophys. Res. Lett.* **33**, L08809 (2006).
- LeBeau, R. P. Jr & Dowling, T. E. EPIC simulations of time-dependent, three-dimensional vortices with application to Neptune's great dark spot. *Icarus* **132**, 239–265 (1998).
- Youssef, A. & Marcus, P. S. The dynamics of jovian white ovals from formation to merger. *Icarus* **162**, 74–93 (2003).
- Fine, K. S., Cass, A. C., Flynn, W. G. & Driscoll, C. F. Relaxation of 2D turbulence to vortex crystals. *Phys. Res. Lett.* **75**, 3277–3280 (1995).
- Schecter, D. A., Dubin, D. H. E., Fine, K. S. & Driscoll, C. F. Vortex crystals from 2D Euler flow: experiment and simulation. *Phys. Fluids* **11**, 905–914 (1999).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The JIRAM project is funded by the Italian Space Agency (ASI). In particular this work has been developed under the ASI-INAF agreement number 2016-23-H.O. The JunoCam instrument and its operations are funded by the National Aeronautics and Space Administration. A portion of this work was supported by NASA funds to the Jet Propulsion Laboratory, to the California Institute of Technology, and to the Southwest Research Institute. A.P.I. was supported by NASA funds to the Juno project and by NSF grant number 1411952.

**Author Contributions** A.A. and C.H. are the Juno mission instrument leads for the JIRAM and JunoCam instruments, respectively, and they planned and implemented the observations discussed in this paper. S.J.B. and J.E.P.C. are respectively the principal and the deputy responsible for the Juno mission. A.A., A. Mura, G.O., J.R., A.I. and F.T.-V. were responsible for writing substantial parts of the paper. M.E.O'N. helped with the interpretation of the cyclonic structure. A. Mura, F.A., M.L.M. and D.G. were responsible for reduction and measurement of the JIRAM data and their rendering into graphical formats. G.E., T.M., G.O. and J.R. were responsible for the same tasks for JunoCam data. F.T.-V. and F.F. were responsible for the geometric calibration of the JIRAM data. G.F., G.S., B.M.D. and S.S. were responsible for the JIRAM data radiance calibrations. A.C., R.N. and R.S. were responsible for the JIRAM ground segment. S.K.A., J.I.L., A. Migliorini, D.T., G.P. and D.T. supervised the work. C.P., A.O. and M.A. were responsible for the JIRAM project from the Italian Space Agency side.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to A.A. ([alberto.adriani@iaps.inaf.it](mailto:alberto.adriani@iaps.inaf.it)).

## METHODS

**The Jovian InfraRed Auroral Mapper.** The Jovian InfraRed Auroral Mapper (JIRAM) is composed of an imager and a spectrometer that share the same telescope<sup>16,17</sup>. The imager focal plane is equipped to observe the planet through a bandpass filter centred at  $4.78\text{ }\mu\text{m}$  with a 480-nm bandwidth (M-band) and a bandpass filter centred at  $3.45\text{ }\mu\text{m}$  with a 290-nm bandwidth (L-band). The spectrometer's slit is optically co-located with the imager's field of view and its spectral range covers the 2–5- $\mu\text{m}$  interval in 336 spectral bins (bands) resulting in a spectral sampling of 8.9 nm per band across the full spectral range. The instrument design allows for acquisition of simultaneous imager and spectrometer observations: in this study, we used the data from the M-band filter, which covers a field of view of about  $1.75^\circ$  by  $6^\circ$  with  $128 \times 432$  pixels. The instantaneous field of view is  $240\text{ }\mu\text{rad}$  (see ref. 16 for instrumental details).

At the time of the observations, the Juno spacecraft was spinning almost perpendicular to the orbital plane. For each spin, JIRAM takes two images: one to the target (nadir direction), and one to the anti-nadir direction, to evaluate the background, which is removed onboard. JIRAM is also equipped with a de-spinning mirror that compensates for the spacecraft rotation and enables it to keep the target image in the field of view during the data acquisition. The de-spinning mirror may also be activated at different times with respect to the nadir direction, allowing a scan of the planet in the spacecraft's spinning plane. No pointing outside the spinning plane is permitted.

The data shown in this paper (integrated radiance from  $4.5\text{ }\mu\text{m}$  to  $5\text{ }\mu\text{m}$ ) have been taken with 12 ms of integration time, resulting in a noise-equivalent radiance lower than  $5 \times 10^{-5}\text{ W sr}^{-1}\text{ m}^{-2}$ . Table 1 shows exact times of observations (start time, stop time and number of observations in that sequence or scan). JIRAM observed both poles with high image quality and spatial resolution during PJ4 and PJ5. Polar coverage during PJ4 was complete for regions within  $30^\circ$  latitude of both poles with a spatial resolution varying between 12 km per pixel and 96 km per pixel for the north pole and between 21 km per pixel and 62 km per pixel for the south pole. JIRAM coverage of the poles during PJ5 was incomplete, limited by JIRAM's field of view and Juno's spin axis orientation at perijove.

For PJ4 at the north pole we have complete data coverage with good emission angle (that is, close to  $90^\circ$ ), but to cover other parts of the planet we also use radiance emitted at lower angles. Such radiance is partially depleted because the absorption due to cold clouds occurs over a longer atmospheric path. A simple correction was applied, mainly with the purpose of better identifying the same features at both PJ4 and PJ5. Since during PJ4 JIRAM observed the same regions of the north pole at different emission angles, those data were used to compile a look-up table that, given radiances measured at emission angles lower than  $90^\circ$ , permits scaling of the measured values to the radiances expected at  $90^\circ$  to make the measurements more comparable to each other.

Data are jovian-located and then re-projected in System III planetocentric geographical coordinates, using a polar orthographic projection. Geometric information was obtained by using *ad hoc* algorithms based on the NAIF-SPICE tool<sup>24</sup> for each image. JIRAM raw data are calibrated in units of radiance ( $\text{W m}^{-2}\text{ sr}^{-1}$ ) as described<sup>16,17</sup>. The responsivity used in this study has been revised to a flat value of  $2 \times 10^6$  digital numbers (DN) per ( $\text{W m}^{-2}\text{ sr}^{-1}$ ) by using the cruise-calibration campaign data, performed by using the orange giant star Aldebaran ( $\alpha$  Tauri) as a reference target.

Finally, the diameters of the cyclones are calculated on the JIRAM infrared images, defining the outer border of the cyclone where the smaller, anticyclonic structures form and planetocentric coordinates are used throughout this report. Images from JIRAM were processed using Matlab (Fig. 1 and Extended Data Fig. 1) and ENVI-IDL (Fig. 3).

Processing of consecutive images allows for animations revealing motion, as well as for quantitative analysis of cloud velocities. JIRAM data in Extended Data Table 1 have been arranged in animations that show the movement of single vortices during PJ4 observations. Each sequence or scan has been composed into a mosaic, and then each mosaic became a frame of the video. We provide nine Supplementary Videos for the north pole (eight for circumpolar cyclones plus the NPC); each video is made of 11 images. We also provide six Supplementary Videos for the south pole (five for circumpolar cyclones plus the SPC); each video is made of 6 images.

**JunoCam.** JunoCam is a visible-spectrum camera designed to acquire images through broadband red, green and blue filters mounted directly on a CCD detector, with an 889-nm methane absorption band filter acquiring an image on a separate rotation typically 30 s later. JunoCam is rigidly mounted on the spinning spacecraft. That way, it uses the spacecraft rotation to take a full panorama within about 30 s consisting of up to 82 narrow exposures, referred to as the 'pushframe' mode. Usually, it takes partial panoramas of the target of interest. The camera has a horizontal field of view of about  $58^\circ$ , and a Kodak KAI-2020 charge-coupled

device (CCD) sensor with four filter stripes, a red, a green, a blue (RGB) and a narrow-band 890-nm infrared filter attached on the  $1,600 \times 1,200$  light-sensitive pixels. For each of the four filters, there is an corresponding readout region of  $1,600 \times 128$  pixels which can be transferred into the resulting raw image. This transfer is not immediate, but the 12-bit data number of each pixel is encoded as an 8-bit value, and tiles of  $16 \times 16$  pixels are compressed in either a lossy or lossless manner. Usually, the encoding of the 12-bit data as an 8-bit value is nonlinear, according to a 'companding' function. Motion blur is mostly avoided by a technique called time delay integration. In colour (RGB) mode, for each exposure, three of the four readout regions are added as stripes to the raw image. Full details of the instrument and its operation are available in ref. 18.

JunoCam observed the same polar regions as JIRAM on PJ4 and PJ5—as well as complementary longitudinal regions on PJ1 and PJ3—but as a visible imager, it acquires images in reflected light. A complete polar view must be pieced together from the unshadowed portions of images collected during multiple perijove passes.

Observations were made in both north and south polar regions during PJ1, PJ3, PJ4 and PJ5. Polar imaging in PJ5 was scheduled over extended periods of time to cover more longitudes as the planet rotates through daylight, which enabled time-lapse measurements that include measurements of rotation of the cyclones.

With an approximate geometrical camera model, including its pointing for each exposure, the appropriate three-dimensional vector was calculated for each pixel in a given reference frame, for example, J2000. Position and pointing information are inferred from SPICE data<sup>24</sup>, with some manual adjustment. Jupiter is modelled as a MacLaurin spheroid on Jupiter's 1-bar level. A planetocentric coordinate system assigns a three-dimensional position to each longitude/latitude pair. The three-dimensional vector, pointing from Juno to the three-dimensional position, completes the connection of each longitude/latitude pair to colour information. With this method, each raw JunoCam image of Jupiter is reduced to an approximately geometrically calibrated polar-map projection.

Because Jupiter is rotating and Juno is moving rapidly, the illumination for each JunoCam image changes rapidly. Comparison of images requires approximate normalization of the images. For now, this is achieved in a heuristic way, essentially stretching contrast over regions of approximately similar solar incidence angles, subtracting the mean brightness for these bins, and accounting for changing light scattering of a presumed haze layer as a function of emission angle, which can be obtained for sufficiently small crops by high-pass filtering. For the JunoCam maps shown in Figs 1 and 2, this correction was made down to a maximum solar illumination angle of  $66^\circ$ , above which the signal-to-noise ratio drops below a value of 3 per pixel. Further nonlinear brightness stretching and saturation enhancement brings out detail.

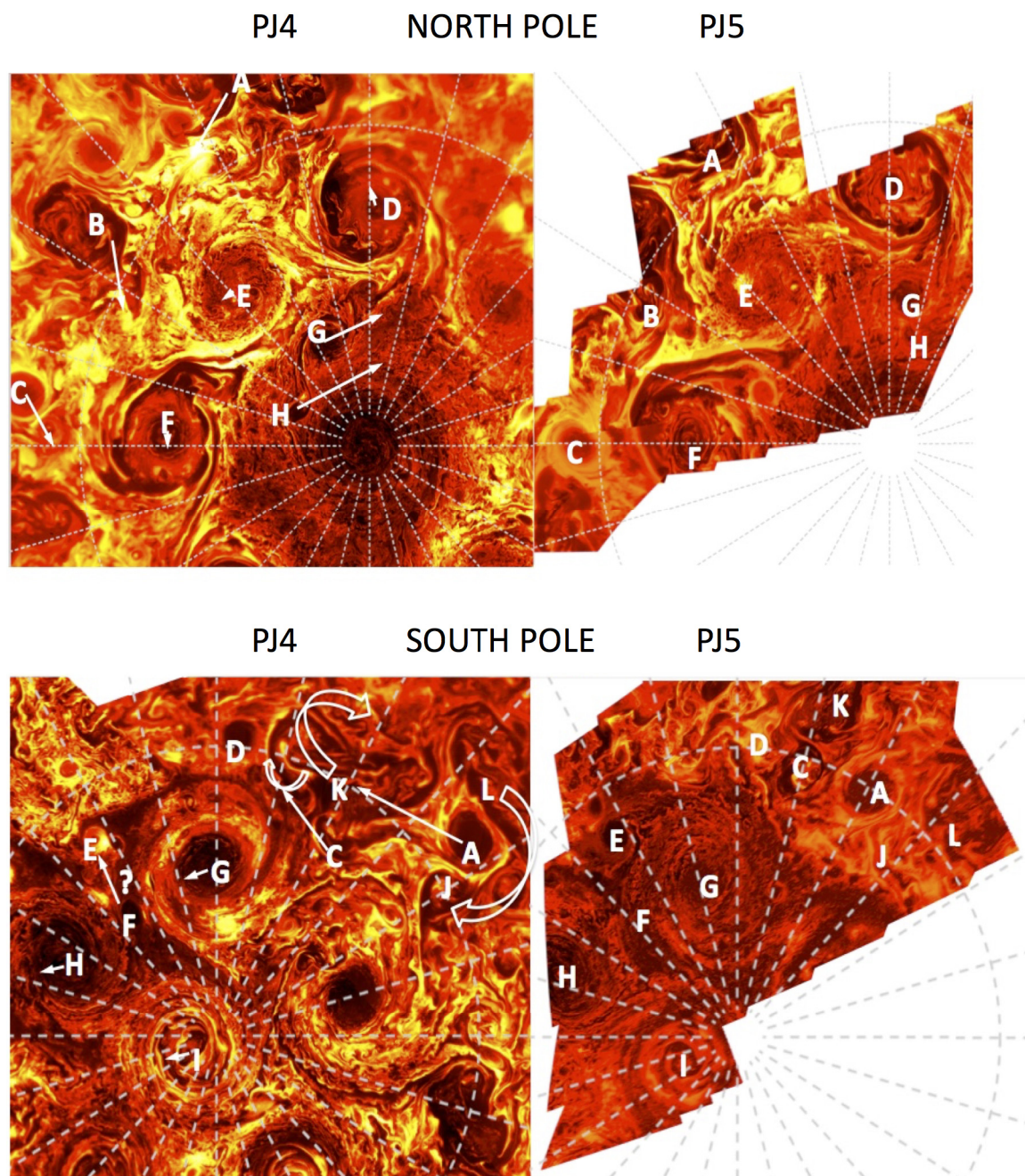
Time sequences of 2 to 3 frames of polar images were made on all perijove passes to track cloud motions. After PJ1, it was clear that the limited time sequence on that orbit verified the visible impression that the features surrounding the poles whose 'arms' implied cyclonic motion really were cyclones. A special effort was made in PJ5 to create longer sequences of time-lapse images that would illustrate subtler motions of the polar features by imaging over a longer time interval. The sequences given below represent three of the best of those animations. A sequence of images from the north pole is available at [http://junocam.pictures/gerald/uploads/20170424/anim/jnc\\_pj05\\_N\\_089\\_to\\_105\\_blend4\\_enh.html](http://junocam.pictures/gerald/uploads/20170424/anim/jnc_pj05_N_089_to_105_blend4_enh.html). A sequence of images from the south pole is available at [http://junocam.pictures/gerald/uploads/20170331/anim/jnc\\_pj05\\_polarS\\_60px\\_lin\\_interpolated\\_21frames\\_1200x1200.html](http://junocam.pictures/gerald/uploads/20170331/anim/jnc_pj05_polarS_60px_lin_interpolated_21frames_1200x1200.html). A close-up version of that south polar sequence is available at [http://junocam.pictures/gerald/uploads/20170406/anim/jnc\\_pj05\\_south\\_polar\\_animation\\_111\\_to\\_121\\_8frames\\_20fps\\_1200px.html](http://junocam.pictures/gerald/uploads/20170406/anim/jnc_pj05_south_polar_animation_111_to_121_8frames_20fps_1200px.html).

**History and terminology of cyclone clusters.** The term 'ditetragonal' has been introduced in the context of crystallography, since it is one of the ten two-dimensional crystallographic point groups (see table 10.1.2.1 in ref. 25). In the non-Euclidean geometry of the curved polar region, a two-dimensional pentagonal rather than a hexagonal pattern would be conceivable, similar to the surface of a pentagon-dodecahedron (see table 10.1.2.2 in ref. 25). But since the size of the vortices does not fit exactly to the geometry of a pentagon-dodecahedron, an unstable structure switching between a hexagon and pentagon could occur, or an oscillating pentagon for vortices of similar sizes. Besides the 'vortex crystals'<sup>22,23</sup> mentioned above, similar vortex patterns also occur in rotating superfluid helium II for quantum-mechanical reasons<sup>26</sup>. Theoretical predictions of such quantized vortices reach back to Onsager<sup>27</sup> and Feynman<sup>28</sup>, although Landau<sup>29</sup> introduced rotons in 1941. The first experimental observations<sup>30</sup> were in 1979.

**Data availability.** The data used for this study will be available once the proprietary period ends, namely about six months after the data were collected by Juno, from the NASA's Planetary Data System at <https://pds.jpl.nasa.gov/tools/data-search/>. The JunoCam data are all available for direct download from the Mission Juno web site in both raw and processed form: <https://www.missionjuno.swri.edu/junocam/processing>.

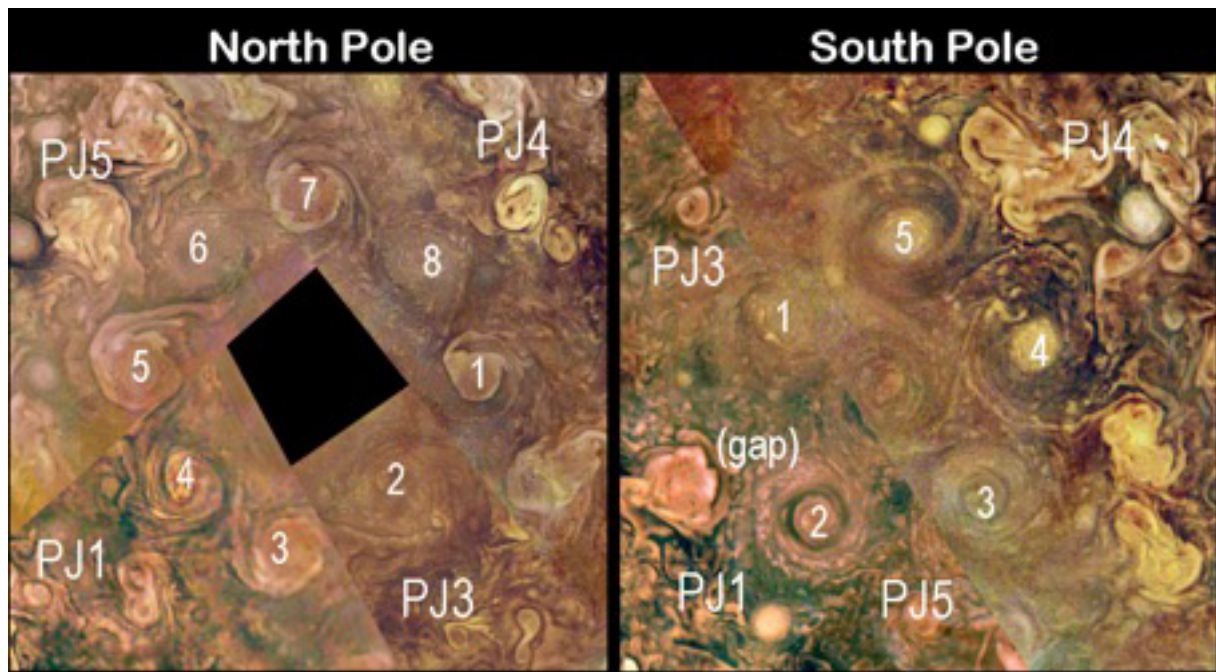
24. Acton, C. H. Ancillary data services of NASA's navigation and ancillary information facility. *Planet. Space Sci.* **44**, 65–70 (1996).
25. Hahn, Th. (ed.) *International Tables for Crystallography* Vol. A, 5th edn, 768, 786 (Springer, 2005).
26. Aref, H. *et al.* Vortex crystals. *Adv. Appl. Mech.* **39**, 1–79 (2003).
27. Gorter, C. J. The two fluid model of helium II. *Nuovo Cimento* **6** (Suppl. 2), 245–250 (1949); see discussion by L. Onsager, 249–250.
28. Feynman, R. P. in *Progress in Low Temperature Physics* Vol. 1 (ed. Gorter, C. J.) Ch. II, 17–53 (Elsevier, 1955).
29. Landau, L. D. The theory of superfluidity of helium II. *Zh. Eksp. Teor. Fiz.* **11**, 592 (1941).
30. Yarmchuk, E. J., Gordon, M. J. V. & Packard, R. E. Observation of stationary vortex arrays in rotating superfluid helium. *Phys. Rev. Lett.* **43**, 214–217 (1979).





**Extended Data Figure 1 | Comparison of the polar cyclonic structures between PJ4 and PJ5.** Here is the comparison between JIRAM 5- $\mu\text{m}$  data acquired during PJ4 and PJ5. The letters identify possible recurrent structures and arrows show the suggested displacements that occurred in the 53-day interval between these two perijoves. The radiance scale is the same as in Fig. 1. When the region surrounding the north pole is not sunlit, there are no JunoCam observations of the NPC. Although the north pole was detected by JIRAM on PJ4, we were unable to determine whether or not it maintains a stable position over the geographic north pole because of insufficient coverage of the NPC during PJ5. However, the cyclonic structures A, B and C move northeast, migrating from the lower latitudes. The G and H internal structures, located between the NPC and the cyclones, are anticyclones and move westward in that narrow corridor between  $85.5^\circ\text{N}$  and  $87^\circ\text{N}$  to their new location observed during PJ5 between vortex D and the NPC. In contrast, JIRAM was able to observe

the SPC in both PJ4 and PJ5. In fact, along with the cyclones G and H shown, the SPC moves northward, increasing its distance with respect to the geographic south pole by  $1.5^\circ$  between PJ4 and PJ5. On the other hand, JunoCam was able to observe the SPC at all perijoves, and found that it was always displaced from the south pole in approximately the same direction (towards a System III longitude of about  $219^\circ \pm 21^\circ$ ), with its central latitude varying from  $88.0^\circ\text{S}$  at PJ1 up to  $89.0^\circ\text{S}$  at PJ4, and down to  $88.4^\circ\text{S}$  at PJ5. It remains to be seen whether this is a cyclic oscillation. The five cyclones remain at almost constant radial distances from the centre of the SPC (and thus not from the geographic south pole), so the whole pentagon drifts in latitude. Anticyclone A appears to move as much as about  $1^\circ$  south and about  $24^\circ$  east. It is forced and surrounded by the two cyclonic structures that consolidate themselves between PJ4 and PJ5 from the origins L, J, C and K. Finally, the anticyclone D disappears while F is expelled from its position and possibly moves to new position E.



**Extended Data Figure 2 | Annotated version of the JunoCam images of the poles.** The unannotated version is shown in Fig. 2. The composite components from each perijove pass that were used to create the figure are noted. Each corresponds to the polar image taken at a time that minimized the emission angle over most of the pole, as detailed in Extended Data Table 2. The PJ4 component is identical to its contribution in Fig. 1, with contributions from the other perijove passes, separated by approximately 90° in longitude, as noted. The northern cyclones forming the inner

square (actually a rhombus) of the ditetragonal pattern are labelled by odd numbers and those forming the outer square by even numbers. The southern cyclones forming a quasi-pentagonal shape are numbered sequentially, with the largest spacing between cyclones labelled 1 and 5, indicated by the 'gap' label. Despite the time differences of 53 to 106 terrestrial days between JIRAM images acquired on PJ4, shown broadly in Fig. 1b and d, and JunoCam images in PJ1, PJ3 and PJ5, the positions of the cyclones are remarkably consistent in System III longitude.



Extended Data Table 1 | JIRAM start time, stop time and number of observations for the different datasets used for this study

North Pole, 4 <sup>th</sup> perijove, 2017-02-02		
Start UTC	Stop UTC	observations
08:10:24	08:15:59	12 <sup>(1)</sup>
08:18:31	08:24:07	12 <sup>(1)</sup>
08:26:08	08:31:44	12 <sup>(1)</sup>
08:34:16	08:39:51	12 <sup>(1)</sup>
08:42:23	08:47:59	12 <sup>(1)</sup>
08:50:31	08:56:06	12 <sup>(1)</sup>
08:58:08	09:03:43	12 <sup>(1)</sup>
09:06:15	09:11:51	12 <sup>(1)</sup>
09:14:23	09:19:58	12 <sup>(1)</sup>
09:22:30	09:28:06	12 <sup>(1)</sup>
09:30:07	09:35:43	12 <sup>(1)</sup>
09:38:15	09:43:50	12 <sup>(1)</sup>
09:46:22	09:51:58	12 <sup>(1)</sup>
11:00:00	11:05:35	12 <sup>(2)</sup>
11:08:07	11:13:43	12 <sup>(2)</sup>
11:16:14	11:21:50	12 <sup>(2)</sup>
11:24:22	11:29:57	12 <sup>(2)</sup>
11:31:59	11:37:34	12 <sup>(2)</sup>
11:40:06	11:45:41	12 <sup>(2)</sup>
11:48:13	11:53:48	12 <sup>(2)</sup>
11:56:20	12:01:55	12 <sup>(2)</sup>
12:04:26	12:10:02	12 <sup>(2)</sup>
12:12:02	12:17:38	12 <sup>(2)</sup>
12:20:07	12:25:44	12 <sup>(2)</sup>

South Pole, 4 <sup>th</sup> perijove, 2017-02-02		
Start UTC	Stop UTC	observations
13:58:59	14:08:08	19
14:18:46	14:27:55	19
14:39:04	14:48:13	19
14:58:53	15:08:02	19
15:18:41	15:27:50	19
15:39:00	15:48:09	19

North Pole, 5 <sup>th</sup> perijove, 2017-03-27		
Start UTC	Stop UTC	observations
08:14:20	08:16:20	5
08:18:49	08:24:19	12
08:26:48	08:30:48	9
08:33:17	08:33:47	2

South Pole, 5 <sup>th</sup> perijove, 2017-03-27		
Start UTC	Stop UTC	observations
09:17:54	09:18:54	3
09:21:23	09:24:24	7
09:26:53	09:29:53	7
09:32:23	09:33:53	4

UTC, coordinated universal time. (1) Approach phase, low resolution; used only to fill small gaps in Fig. 1; (2) minimum emission angle, high resolution, used to make most of the mosaic in Fig. 1.



Extended Data Table 2 | Details of the JunoCam observations

Pole	Perijove	Date	Time (UTC)	File Name
North	1	2016-08-27	11:57:40- 11:57:49	JNCE_2016240_00C06160_V02_553
South	1	2016-08-27	11:59:12- 11:59:21	JNCE_2016240_00C06186_V02_579
North	3	2016-12-11	16:13:45- 16:13:54	JNCE_2016346_03C00099_V02_588
South	3	2016-12-11	18:10:49- 18:10:56	JNCE_2016346_03C00126_V02_615
North	4	2017-02-02	12:08:22- 12:08:32	JNCE_2017033_04C00097_V01_624
South	4	2017-02-02	14:06:29- 14:06:38	JNCE_2017033_04C00109_V01_636
North	5	2017-05-19	08:04:43- 08:04:52	JNCE_2017086_05C00102_V01_890
South	5	2017-05-19	10:01:48- 10:01:55	JNCE_2017086_05C00118_V01_901

The observations listed correspond to those used in Figs 1 and 2 and in Extended Data Fig. 2.

# Measurement of Jupiter's asymmetric gravity field

L. Iess<sup>1</sup>, W. M. Folkner<sup>2</sup>, D. Durante<sup>1</sup>, M. Parisi<sup>2</sup>, Y. Kaspi<sup>3</sup>, E. Galanti<sup>3</sup>, T. Guillot<sup>4</sup>, W. B. Hubbard<sup>5</sup>, D. J. Stevenson<sup>6</sup>, J. D. Anderson<sup>7</sup>, D. R. Buccino<sup>2</sup>, L. Gomez Casajus<sup>8</sup>, A. Milani<sup>9</sup>, R. Park<sup>2</sup>, P. Racioppa<sup>1</sup>, D. Serra<sup>9</sup>, P. Tortora<sup>8</sup>, M. Zannoni<sup>8</sup>, H. Cao<sup>6</sup>, R. Helled<sup>10</sup>, J. I. Lunine<sup>11</sup>, Y. Miguel<sup>4</sup>, B. Militzer<sup>12</sup>, S. Wahl<sup>12</sup>, J. E. P. Connerney<sup>13</sup>, S. M. Levin<sup>2</sup> & S. J. Bolton<sup>7</sup>

**The gravity harmonics of a fluid, rotating planet can be decomposed into static components arising from solid-body rotation and dynamic components arising from flows. In the absence of internal dynamics, the gravity field is axially and hemispherically symmetric and is dominated by even zonal gravity harmonics  $J_{2n}$  that are approximately proportional to  $q^n$ , where  $q$  is the ratio between centrifugal acceleration and gravity at the planet's equator<sup>1</sup>. Any asymmetry in the gravity field is attributed to differential rotation and deep atmospheric flows. The odd harmonics,  $J_3$ ,  $J_5$ ,  $J_7$ ,  $J_9$  and higher, are a measure of the depth of the winds in the different zones of the atmosphere<sup>2,3</sup>. Here we report measurements of Jupiter's gravity harmonics (both even and odd) through precise Doppler tracking of the Juno spacecraft in its polar orbit around Jupiter. We find a north–south asymmetry, which is a signature of atmospheric and interior flows. Analysis of the harmonics, described in two accompanying papers<sup>4,5</sup>, provides the vertical profile of the winds and precise constraints for the depth of Jupiter's dynamical atmosphere.**

The external, harmonic, gravitational potential of a body can be expanded into a series of complex spherical harmonic functions  $Y_{lm}(\theta, \phi)$  (an orthonormal basis for functions defined on the unit sphere, with each element defined by its degree  $l$  and order  $m$ ), multiplied by a scaling factor that depends on the normalized radial distance  $r/R$  from the centre of the body:

$$U(r, \theta, \varphi) = -\frac{GM}{r} \left[ 1 + \sum_{l \geq 2} \left( \frac{R}{r} \right)^l \sum_{m=-l}^l U_{lm} Y_{lm}(\theta, \varphi) \right]$$

where  $GM$  is the gravitational parameter. For a planet,  $R$  is generally chosen as the equatorial radius of the body. Were the internal density  $\rho$  of the body known, the harmonic coefficients  $U_{lm}$  could be obtained from the integral over the volume  $V$  of the body (see ref. 6 and references therein):

$$U_{lm} = \frac{1}{(2l+1)MR^l} \int_V (r')^l Y_{lm}^*(\theta', \varphi') \rho(r', \theta', \varphi') dV'$$

When the density does not depend on longitude, as expected for a fluid and rapidly rotating planet such as Jupiter, the above expression can be simplified:

$$J_l = -\sqrt{2l+1} U_{l0} = -\frac{1}{MR^l} \int_V (r')^l P_l(\theta') \rho(r', \theta') dV'$$

where  $P_l$  is the Legendre polynomial of degree  $l$ . Thus, the zonal coefficients  $J_l$  bear important, although non-unique, information about the density distribution inside Jupiter.

On 4 July 2016, the Juno spacecraft was captured by the gravity field of Jupiter, starting its prime mission—the investigation of the deep interior, the magnetosphere and the atmosphere of the planet. The

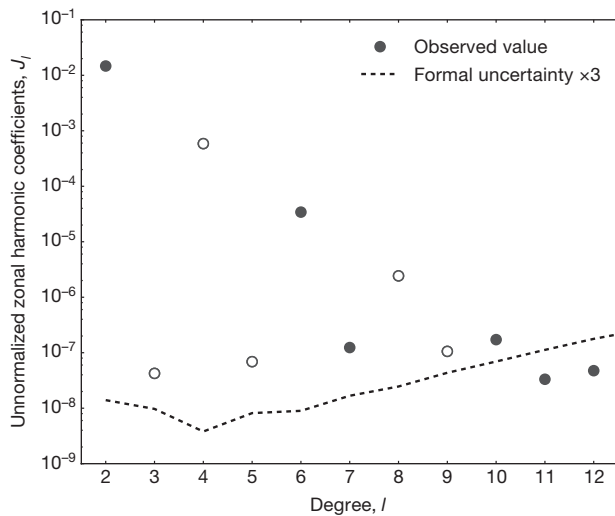
spacecraft is currently in a highly eccentric (eccentricity  $e = 0.98$ ), long-period (52.9 days) polar orbit, with a pericentre altitude of about 4,000 km above the 1-bar level, as inferred from radio occultations<sup>7</sup>.

As a consequence of the equivalence principle, gravity field determinations require the measurement of the relative motion of (at least) two masses. In the Juno gravity experiment, the spacecraft acts as a test particle falling in the gravity field of the planet. Earth is the second end mass. Jupiter's gravity is inferred from range-rate measurements between a ground antenna and the spacecraft during perijove passes. To measure Jupiter's gravity field, the ground station transmits two carrier signals, at 7,153 MHz (X band) and 34,315 MHz (Ka band). On board, an X-band transponder and a dedicated Ka-band frequency translator (a radio-science instrument) lock the incoming carrier signals and retransmit them back to the ground station at 8,404 MHz and 32,088 MHz, respectively. The range-rate (Doppler) observable is obtained by comparing the transmitted and received frequencies. Juno is the first deep-space mission that uses Ka-band radio systems for planetary geodesy. The Ka-band and multi-frequency radio links have previously been used only for precision tests of relativistic gravity in the cruise phase of the Cassini spacecraft<sup>8,9</sup>. Owing to the dispersion properties of plasmas, Ka-band radio links provide excellent immunity to the adverse effects of charged particles along the propagation path, including the Io torus (a potential source of bias in the gravity estimates; see Methods). The Juno radio system enables further reduction of plasma noise (an additional approximately 75%) by combining X- and Ka-band Doppler observables<sup>10</sup>. To reduce the noise from tropospheric water vapour, a radiometer placed near the ground antenna continuously monitors the wet path delay along the line of sight.

Our analysis is based on the first two Ka-band perijove passes of Juno, labelled PJ3 (11 December 2016) and PJ6 (19 May 2017). Doppler measurements were integrated over 60 s before processing to enable adequate sampling of the gravity signal. At this timescale, the measured two-way range-rate noise in the Ka band was  $1.5 \times 10^{-5} \text{ m s}^{-1}$  for an integration time of 60 s, in line with the expectations from Ka-band radio link noise models<sup>11</sup>. The Doppler noise is approximately white between  $4 \times 10^{-4} \text{ Hz}$  and  $2 \times 10^{-2} \text{ Hz}$  (the characteristic frequency range of the gravity signal).

The dynamical model used in the orbital fit is driven by the theoretical expectations for the gravity field of gaseous planets. We adopt here the standard spherical harmonics representation of planetary gravity fields, whose expansion coefficients are determined by the density distribution inside the body (ref. 6 and references therein). Models of Jupiter's interior structure predict that the planet's gravity is dominated by an axially and hemispherically symmetric component attributed to solid-body rotation<sup>12,13</sup>. This component is determined by the radial density distribution in the rotating planet and is represented by even zonal harmonic coefficients  $J_{2n} \propto q^n$ . Atmospheric and internal dynamics can produce small density perturbations that result in a more complex gravity representation, involving

<sup>1</sup>Sapienza Università di Roma, 00184 Rome, Italy. <sup>2</sup>Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. <sup>3</sup>Weizmann Institute of Science, Rehovot 76100, Israel. <sup>4</sup>Observatoire de la Côte d'Azur, 06304 Nice, France. <sup>5</sup>Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA. <sup>6</sup>California Institute of Technology, Pasadena, California 91125, USA. <sup>7</sup>Southwest Research Institute, San Antonio, Texas 78238, USA. <sup>8</sup>Università di Bologna, 47100 Forlì, Italy. <sup>9</sup>Università di Pisa, 56127 Pisa, Italy. <sup>10</sup>University of Zurich, 8057 Zurich, Switzerland. <sup>11</sup>Cornell University, Ithaca, New York 14853, USA. <sup>12</sup>University of California, Berkeley, California 94720, USA. <sup>13</sup>NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA.



**Figure 1 | Zonal gravity harmonic coefficients  $J_2$ – $J_{12}$ .** The dashed line shows the realistic uncertainty (Table 1). Solid and empty circles denote positive and negative values, respectively.

odd zonal and possibly tesseral harmonics, as well as small corrections to the even zonal harmonics<sup>3,5,14</sup>. The latter are however indiscernible from the much larger contribution of solid-body rotation up to harmonics of degree 12, where the dynamics is expected to dominate the gravity signal<sup>2</sup>. Hence, any detection of an asymmetric (hemispherically or axially) gravity field would be a signature of internal dynamics due to flows. Juno tracking data have provided evidence of hemispherical (north–south) asymmetries in the gravity field of a giant planet.

Prior to PJ3, the best estimate of Jupiter's even zonal gravity field was obtained using noisier X-band Doppler observables from the first two Juno perijove passes (PJ1 and PJ2)<sup>15,16</sup>. These early results improved previous estimates<sup>17,18</sup> of the zonal harmonic coefficients  $J_4$  and  $J_6$  and allowed the determination of  $J_8$ . Those measurements of  $J_4$  and  $J_6$  have been used to constrain the radial density profile of the planet<sup>19</sup>. However, the magnitude of the much smaller odd zonal field could not be determined, because of the unfavourable observation geometry and the large propagation noise caused by interplanetary plasma on the X-band uplink signal (7.2 GHz).

High-accuracy Ka-band data acquired during PJ3 and PJ6 provided the first estimate of the asymmetric component of Jupiter's gravity (Fig. 1 and Table 1). We processed Doppler data using orbit determination codes developed for spacecraft navigation (the MONTE software of the Jet Propulsion Laboratory) and an external estimation filter. Data from PJ3 and PJ6 were separately fitted with the spacecraft state vector at the beginning of the tracking pass (about 6 h before transit at the pericentre), Jupiter's gravitational parameter  $GM$ , the zonal harmonic coefficients  $J_2$ – $J_{24}$ , the tesseral quadrupole harmonics, the pole position and rate at epoch J2017.0 (1 January 2017, 12:00 UTC) and the

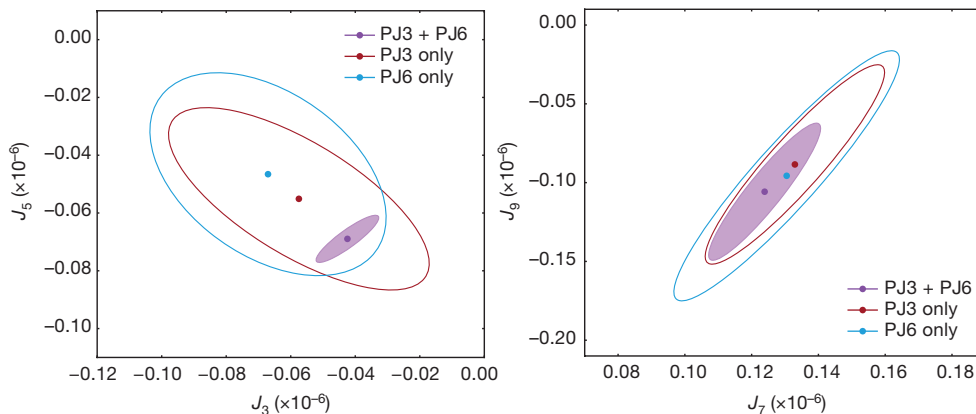
**Table 1 | Gravity solution**

	Value	Uncertainty
$J_2 (\times 10^{-6})$	14,696.572	0.014
$C_{21} (\times 10^{-6})$	−0.013	0.015
$S_{21} (\times 10^{-6})$	−0.003	0.026
$C_{22} (\times 10^{-6})$	0.000	0.008
$S_{22} (\times 10^{-6})$	0.000	0.011
$J_3 (\times 10^{-6})$	−0.042	0.010
$J_4 (\times 10^{-6})$	−586.609	0.004
$J_5 (\times 10^{-6})$	−0.069	0.008
$J_6 (\times 10^{-6})$	34.198	0.009
$J_7 (\times 10^{-6})$	0.124	0.017
$J_8 (\times 10^{-6})$	−2.426	0.025
$J_9 (\times 10^{-6})$	−0.106	0.044
$J_{10} (\times 10^{-6})$	0.172	0.069
$J_{11} (\times 10^{-6})$	0.033	0.112
$J_{12} (\times 10^{-6})$	0.047	0.178
$k_{22}$	0.625	0.063
$\alpha (^\circ)$	268.0570	0.0013
$\delta (^\circ)$	64.4973	0.0014

Jupiter's gravity harmonics coefficients (unnormalized; reference radius 71,492 km), the Love number  $k_{22}$  and the pole coordinates ( $\alpha$ , right ascension;  $\delta$ , declination) at epoch J2017.0, obtained from the PJ3 and PJ6 Juno science orbits. The deviation of the principal axis of inertia from the spin axis, as inferred from the uncertainty in  $C_{21} = \text{Re}(U_{21})\sqrt{5/3}$  and  $S_{21} = \text{Im}(U_{21})\sqrt{5/3}$ , is smaller than about 0.4 arcsec (130 m at the reference radius).  $J_2$  includes a tidal term currently estimated at about  $2.98 \times 10^{-8}$ . The associated uncertainties are realistic values that can be used for analysis and interpretation and correspond to three times the formal  $1\sigma$  uncertainties.

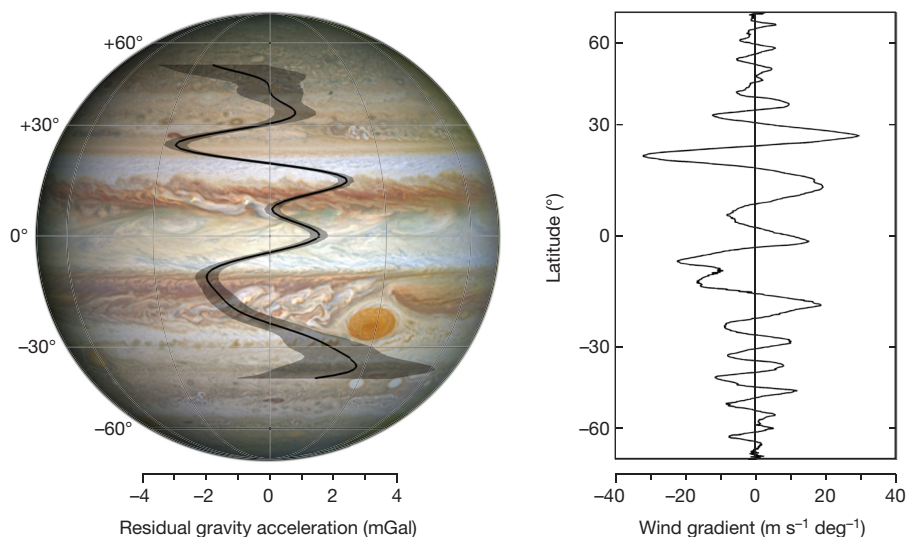
$k_{22}$  Love number. This set of parameters allows the fitting of all data to the noise level. The  $l=2$  tesseral coefficients, although not strictly required for a least-size solution, were estimated to search for a possible deviation of the principal axis of inertia from the spin axis. We adopted the masses and the ephemerides of the Jovian satellites from file<sup>18</sup> jup310 of NASA's Navigation and Ancillary Information Facility and considered their uncertainties in the final covariance matrix. A linear correction to Jupiter's orbit was applied to fit range data acquired in the X band during the tracking pass. The relativistic Lense–Thirring precession was included and the magnitude of Jupiter's polar moment of inertia was set to interior model predictions, considered with 20% of uncertainty (affecting the recovery of Jupiter's spin axis). The single-arc solutions were then combined in a global multi-arc solution made up by two categories of parameters: local (pertaining to each arc) and global (common to both arcs). Only the initial spacecraft conditions were treated as local parameters. No constraints were applied to the global parameters, except Jupiter's gravitational parameter, whose current estimate is more accurate than that obtained so far from Juno (see Methods). The data were weighted according to the Doppler noise in each Ka-band pass, assuming no correlation between samples. The correctness of this assumption was verified a posteriori from the nearly white power spectral density of the residuals in the frequency band of interest (see Methods).

The two single-arc gravity solutions are fully compatible at  $2\sigma$ , except for  $J_4$  ( $3.5\sigma$ ; see Fig. 2 for examples). Fitting PJ3 and PJ6 data jointly does



**Figure 2 |  $3\sigma$  uncertainty ellipses of  $J_3$ – $J_5$  and  $J_7$ – $J_9$ .** Brown and cyan ellipses represent single-arc PJ3 and PJ6 solutions, respectively. The solid violet ellipse shows the PJ3 + PJ6 combined solution.





**Figure 3 | Gravity disturbances due to atmospheric dynamics.** **a**, An image of Jupiter taken by the Hubble Wide Field Camera in 2014 (<https://en.wikipedia.org/wiki/Jupiter>), showing the latitudinal dependence of residual gravity acceleration (in milligals, positive outwards) and associated  $3\sigma$  uncertainty (shaded area) at a reference distance of 71,492 km, when the gravity from the even zonal harmonics  $J_2$ ,  $J_4$ ,  $J_6$  and  $J_8$  is removed. The residual gravity field, which is dominated by the dynamics of the flows, shows marked peaks correlated with the band structure. **b**, Latitudinal gradient of the measured wind profile. The largest (negative) peak of  $-3.4 \pm 0.4$  mGal ( $3\sigma$ ) is found at a latitude of  $24^\circ$  N, where the latitudinal gradient of the wind speed reaches its largest value. The relation between the gravity disturbances and wind gradients is discussed in an accompanying paper<sup>4</sup>.

not require any tesseral components other than the quadrupole, even if the two ground tracks are separated by about  $150^\circ$ . However, the available data do not allow us to set a reliable upper limit to tesseral harmonics, although numerical simulations indicate that a tesseral field corresponding to a flow depth larger than 380 km would produce signatures in the Doppler residuals (see Methods and ref. 20). The considered covariances that correspond to this flow depth are larger than the uncertainties reported in Table 1. The current dataset does not show evidence of a time-varying gravity field, as may result from Jupiter's normal modes<sup>21</sup>.

For large atmospheric flows on rotating planets, wind shear is accompanied by density gradients; therefore, it is possible to link the flows and the gravity field directly. The velocity gradient affects both the even and odd zonal harmonic coefficients, but only the odd coefficients bear the unique signature of the dynamics when  $l < 10$  (for  $l > 10$  the even coefficients are also dominated by the dynamics of the flows; see Fig. 1). We singled out the contribution of the winds by removing the  $J_2$ ,  $J_4$ ,  $J_6$  and  $J_8$  harmonic components from the complete gravity potential. The north-south asymmetry component of gravity acceleration reaches its largest magnitude of  $3.4 \pm 0.4$  mGal ( $3\sigma$ ;  $1 \text{ Gal} = 1 \text{ cm s}^{-2}$ ) at a latitude of  $24^\circ$  N, approximately at the transition between the northern equatorial belt and the northern tropical zone (Fig. 3). Remarkably, this region corresponds to a large velocity and latitudinal gradient of surface winds, as expected for a gravity signal, owing to wind dynamics<sup>4,14</sup>. The odd zonal harmonics  $J_3$ ,  $J_5$ ,  $J_7$  and  $J_9$  and the associated gravity acceleration may be used to infer the depth and the vertical profile of the winds<sup>3,4</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 20 September 2017; accepted 18 January 2018.**

- Lanzano, P. The equilibrium of a rotating body of arbitrary density. *Astrophys. Space Sci.* **29**, 161–178 (1974).
- Hubbard, W. B. Gravitational signature of Jupiter's deep zonal flows. *Icarus* **137**, 357–359 (1999).
- Kaspi, Y. Inferring the depth of the zonal jets on Jupiter and Saturn from odd gravity harmonics. *Geophys. Res. Lett.* **40**, 676–680 (2013).
- Kaspi, Y. et al. Jupiter's atmospheric jet streams extend thousands of kilometres deep. *Nature* **555**, <https://doi.org/10.1038/nature25793> (2018).
- Guillot, T. et al. A suppression of differential rotation in Jupiter's deep interior. *Nature* **555**, <https://doi.org/10.1038/nature25775> (2018).
- Bertotti, B., Farinella, P. & Vokrouhlický, D. *Physics of the Solar System: Dynamics and Evolution, Space Physics, and Spacetime Structure* 35–61 (Springer, 2012).
- Lindal, G. F. et al. The atmosphere of Jupiter: an analysis of the Voyager radio occultation measurements. *J. Geophys. Res.* **86**, 8721–8727 (1981).
- Bertotti, B., Iess, L. & Tortora, P. A test of general relativity using radio links with the Cassini spacecraft. *Nature* **425**, 374–376 (2003).
- Armstrong, J. W., Iess, L., Tortora, P. & Bertotti, B. Stochastic gravitational wave background: upper limits in the  $10^{-6}$  to  $10^{-3}$  Hz band. *Astrophys. J.* **599**, 806–813 (2003).

- Bertotti, B., Comoretto, G. & Iess, L. Doppler tracking of spacecraft with multifrequency links. *Astron. Astrophys.* **269**, 608–616 (1993).
- Asmar, S. W., Armstrong, J. W., Iess, L. & Tortora, P. Spacecraft Doppler tracking: noise budget and accuracy achievable in precision radio science observations. *Radio Sci.* **40**, RS2001 (2004).
- Hubbard, W. B. Effects of differential rotation on the gravitational figures of Jupiter and Saturn. *Icarus* **52**, 509–515 (1982).
- Hubbard, W. B. & Militzer, B. A preliminary Jupiter model. *Astrophys. J.* **820**, 80 (2016).
- Kaspi, Y., Hubbard, W. B., Showman, A. P. & Flierl, G. R. Gravitational signature of Jupiter's internal dynamics. *Geophys. Res. Lett.* **37**, L01204 (2010).
- Bolton, S. J. et al. Jupiter's interior and deep atmosphere: the initial pole-to-pole passes with the Juno spacecraft. *Science* **356**, 821–825 (2017).
- Folkner, W. M. et al. Jupiter gravity field estimated from the first two Juno orbits. *Geophys. Res. Lett.* **44**, 4694–4700 (2017).
- Jacobson, R., Haw, R., McElrath, T. & Antreasian, P. A comprehensive orbit reconstruction for the Galileo prime mission. *Adv. Astronaut. Sci.* **103**, 465–486 (1999).
- Jacobson, R. A. *Jupiter satellite ephemeris file jup310.bsp* [https://naif.jpl.nasa.gov/pub/naif/generic\\_kernels/spk/satellites/](https://naif.jpl.nasa.gov/pub/naif/generic_kernels/spk/satellites/) (2009).
- Wahl, S. M. et al. Comparing Jupiter interior structure models to Juno gravity measurements and the role of an expanded core. *Geophys. Res. Lett.* **44**, 4649–4659 (2017).
- Parisi, M. et al. Probing the depth of Jupiter's Great Red Spot with the Juno gravity experiment. *Icarus* **267**, 232–242 (2016).
- Durante, D., Guillot, T. & Iess, L. The effect of Jupiter oscillations on Juno gravity measurements. *Icarus* **282**, 174–182 (2017).

**Acknowledgements** This research was carried out at the Sapienza University of Rome, University of Bologna and University of Pisa under the sponsorship of the Italian Space Agency; at the Jet Propulsion Laboratory, California Institute of Technology under a NASA contract; by the Southwest Research Institute under a NASA contract. Support was provided also by the Israeli Space Agency (Y.K. and E.G.) and the Centre National d'Études Spatiales (T.G. and Y.M.). All authors acknowledge support from the Juno Project.

**Author Contributions** L.I. and W.M.F. led the experiment and supervised the data analysis. L.I. wrote most of the manuscript. D.D. and M.P. carried out the gravity data analysis. Y.K. and E.G. provided models of the asymmetric and tesseral gravity field. Y.K., E.G., T.G., W.B.H. and D.J.S. carried out consistency checks with interior models and provided theoretical support. D.R.B. planned and supervised the data collection. P.R. designed and coded the orbit determination filter used in this analysis. L.G.C., P.T. and M.Z. provided the media calibrations. J.D.A., A.M., R.P. and D.S. advised on the data analysis. H.C., R.H., J.I.L., Y.M., B.M. and S.W. helped in the definition of the scientific objectives of the measurements. J.E.P.C., S.M.L. and S.J.B. supervised the planning and execution of the gravity experiment.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to L.I. ([luciano.iess@uniroma1.it](mailto:luciano.iess@uniroma1.it)).

**Reviewer Information** Nature thanks J. Fortney and N. Nettelmann for their contribution to the peer review of this work.

## METHODS

**Data acquisition.** Previous determinations of the Jovian gravity with Juno were carried out using the standard radio system of the spacecraft in the X band (7.2–8.4 GHz) during the first two perijove passes (PJ1 and PJ2). At these lower frequencies, Doppler data were marred by interplanetary plasma noise (although mechanical noise from the ground antenna was considerable in PJ1). Our analysis is based on radio tracking of Juno in the Ka band during two perijove transits on 11 December 2016 (PJ3 at 17:03:40 UTC; UTC, coordinated universal time) and 19 May 2017 (PJ6 at 06:00:45 UTC). The use of the Ka band provided excellent immunity to propagation noises due to charged particles from solar winds and Earth's ionosphere. PJ3 and PJ6 were the first two perijove passes of the mission that were devoted to gravity science. Ground support was provided by DSS 25 (Goldstone, California), the only antenna of NASA's Deep Space Network (DSN) with two-way Ka-band capabilities. Two-way Ka- and X-band data were acquired from 12:47 UTC to 19:19 UTC during PJ3 (about 390 Doppler observables with 60 s integration time for each band), and from 01:39 UTC to 09:25 UTC (about 460 Doppler observables per band) in PJ6. To improve the estimate of the spacecraft trajectory, we also used data acquired in the X band from an antenna of the Canberra DSN complex (DSS 43) after the end of the DSS 25 pass, before an orbit-trimming manoeuvre.

Doppler data were obtained from a wide-band open-loop receiver used for radio-science investigations. A specially designed digital phase-locked loop was applied to the 1-kHz complex samples of the received electric field to obtain the phase history and the sky frequencies. Doppler data from the standard closed-loop receiver are generally noisier, resulting in larger formal uncertainties. The central values of the estimates from the two datasets are statistically compatible.

**Non-gravitational accelerations.** The dynamical model used in the fit is purely deterministic. All non-gravitational forces acting on the spacecraft are modelled using a suitable set of parameters, whose uncertainties contribute to the final covariance matrix. The largest non-gravitational acceleration originates from the solar radiation pressure (about  $9 \times 10^{-9} \text{ m s}^{-2}$ ) acting on the 61-m<sup>2</sup> solar panels and the 3-m high-gain antenna. Modelling this acceleration is simple, as the Sun aspect angle—and therefore the acceleration—is constant during the pass. We have assumed that the reflectivity of the surfaces is known with a 20% uncertainty. Our dynamical model includes the small acceleration from the latitudinally varying, Jovian infrared emission ( $1.2 \times 10^{-9} \text{ m s}^{-2}$  at the equator) and the radiation pressure from the albedo of the planet ( $6 \times 10^{-10} \text{ m s}^{-2}$ ). The negligible effect of inaccurate modelling of these non-gravitational accelerations on the gravity estimate was assessed using numerical simulations. The anisotropic thermal emission from the spacecraft and from possible gas leaks may produce small additional accelerations along the direction of the spin axis (the other components are averaged out). As the direction of Earth and the Sun differ by only 9° during the observations, these accelerations can be confused with the solar radiation pressure, and their effect on the gravity estimate is accounted for in the 20% uncertainty attributed to the solar radiation pressure. Other accelerations, such as atmospheric and magnetic drag, are too small to affect the gravity estimate.

**Orbit geometry.** The orbit geometry is a crucial factor in gravity determinations. The key parameters are the orbital altitude and the angle between the line of sight and the spacecraft acceleration. Juno's pericentre altitudes are sufficiently low (4,154 km in PJ3 and 3,503 km in PJ6) to reveal density inhomogeneities with spatial scales much smaller than the radius of the planet. On the other hand, the large eccentricity causes the radial distance from the planet to increase quickly with latitude, strongly reducing the sensitivity to gravity disturbances in the polar regions (more markedly in the southern hemisphere, owing to the location of the pericentre north of the equator). The eccentricity of the orbit limits also the gravitational contact time: the spacecraft covers 60° in latitude in about 1,200 s, reaching a velocity of about 60 km s<sup>-1</sup> at the pericentre. The other factor that affects the recovery of the gravity field is the orientation of the orbital plane with respect to Earth, which controls the projection of the spacecraft velocity along the line of sight. Although the angle between the negative orbit normal and Earth's direction is not optimal (19.2° in PJ3 and 15.1° in PJ6), the projected velocity and acceleration still provide good observability of the zonal field.

Owing to Jupiter's oblateness, the pericentre drifts northward by about 1° per orbit from an initial latitude of 2.7°. At the end of the nominal mission, it will reach a latitude of 32.6° N, allowing a better determination of gravity at high northern latitudes. The node longitude is controlled by orbital manoeuvres to target specific Jupiter longitudes and obtain uniform coverage of the planet's surface. These manoeuvres are carried out far from the pericentre and therefore do not affect the gravity determinations. The orientation of the orbital plane with respect to Earth changes from a nearly face-on configuration at orbit insertion to edge-on after about three years. Detailed information on Juno's orbit can be obtained from NASA's HORIZONS system (<https://ssd.jpl.nasa.gov>). Extended Data Table 1 reports the main geometrical parameters that are relevant to gravity determination.

**Data quality and calibration.** We have carefully assessed and ruled out considerable biases in the gravity estimate due to systematic effects in the data and the dynamical model. The largest systematic effect in the Doppler measurement is from the dry troposphere, which causes path delay variations up to about  $3 \times 10^{-4} \text{ m s}^{-1}$  over timescales of 6–8 h. The suppression of this large signal is obtained using ground meteorological data (mostly surface pressure and temperature) and careful modelling of elevation-dependent effects. Although a small residual tropospheric signal (mostly due to horizontal pressure gradients) cannot be excluded, its timescale is much larger than that of the gravity harmonics (10–30 min). Its effect on the gravity determination is therefore negligible.

The path delay due to the ionospheric plasma is strongly reduced thanks to the use of Ka band. The DSN provides calibrations of the ionospheric path delays at each tracking complex by mapping dual-frequency GPS (Global Positioning System) measurements onto the line of sight of the spacecraft. The applied corrections never exceed a few centimetres over timescales of several hours, corresponding to path delay rates of about  $2 \times 10^{-6} \text{ m s}^{-1}$ . Although inherently small, these effects can be further reduced thanks to GPS-based calibrations.

According to models of Doppler noise in Ka-band interplanetary radio links<sup>11</sup>, solar wind turbulence becomes a dominant noise source only at solar elongation angles lower than 15° when partial calibration aided by the X-band radio link is available<sup>10,22</sup>. For Juno, the expected interplanetary plasma noise in PJ3 (elongation 61.6°) and PJ6 (elongation 135.4°) is  $3 \times 10^{-7} \text{ m s}^{-1}$  and  $1 \times 10^{-7} \text{ m s}^{-1}$ , respectively, with 60 s integration times. These values are well below the contributions expected from the wet-troposphere path delay variation and mechanical noise from the antenna<sup>11</sup>. Path delay variations due to tropospheric water vapour were calibrated using two microwave radiometers located near the ground antenna, with parallel lines of sight. After these calibrations, Doppler noise at 60-s integration time decreased by about 30%.

The timescale of gravity measurements is determined by the spatial scale of the gravity field and by the spacecraft velocity. For the gravity harmonic of degree  $l$ , the timescale is roughly  $\pi R / V_{sc}$ , where  $R$  is Jupiter's equatorial radius and  $V_{sc}$  is the velocity of the spacecraft near the pericentre. For  $l = 12$ , the timescale of the gravity signal is about 300 s. The Doppler measurements were integrated over 60 s before processing to ensure adequate sampling of the gravity signal. At this timescale, the measured range-rate noise in the Ka band was  $1.5 \times 10^{-5} \text{ m s}^{-1}$  at 60 s, in line with the estimates of Ka-band radio link noise models<sup>11</sup>. The PJ3 and PJ6 Doppler residuals after plasma and tropospheric calibrations and the corresponding Allan deviations are shown in Extended Data Figs 1 and 2. The slope of the Allan deviation (approximately proportional to the inverse square root of the integration time) is consistent with a white Doppler noise between  $4 \times 10^{-4}$  and  $2 \times 10^{-2} \text{ Hz}$  (the band of the gravity signal). The low Doppler noise experienced by Juno is much smaller than the gravity signal from the odd harmonics (an example is shown in Extended Data Fig. 3), facilitating their identification.

**Effect of the Io plasma torus.** Juno's radio signal invariably crosses the region of charged particles generated by the ionization of the gases emitted by Io's volcanoes, known as the Io torus. The resulting path delay variation may be an important source of bias in the gravity estimates. The plasma density of the Io torus shows a variability of a factor of 2 over time periods of around 20 days and is difficult to model<sup>23</sup>. The path delay variation during a Juno pass can be estimated and partially calibrated by means of differential Doppler measurements in the X and Ka bands. In PJ3 and PJ6, we measured path delay variations ascribed to the Io torus of about 2–4 cm in the Ka band (16 times larger in the X band) over a time period of about two hours.

The fractional frequency shift  $y$  of the received signal can be modelled as the sum of a non-dispersive contribution  $y_{nd}$  (dominated by the orbital dynamics) and a dispersive contribution from charged particles:

$$y = y_{nd} + k \left( \frac{\dot{P}_u}{f_u^2} + \frac{\dot{P}_d}{\alpha^2 f_u^2} + \frac{\dot{I}_u}{f_u^2} + \frac{\dot{I}_d}{\alpha^2 f_u^2} \right) \quad (1)$$

Here  $f_u$  is the frequency of the signal transmitted by the ground station and  $\alpha$  is the transponding ratio (the ratio between the frequencies transmitted and received by the spacecraft).  $\dot{P}_u$ ,  $\dot{P}_d$ ,  $\dot{I}_u$  and  $\dot{I}_d$  are the time derivatives of the columnar electron content from the interplanetary and ionospheric plasma and the Io torus, respectively, in the uplink (subscript 'u') and downlink ('d') paths. The constant  $k = e^2 / (8\pi^2 \epsilon_0 m_e c)$  is approximately  $1.34 \times 10^{-7} \text{ m}^2 \text{ s}^{-1}$ , where  $e$  and  $m_e$  are the charge and mass of the electron, respectively,  $\epsilon_0$  is the vacuum permittivity and  $c$  is the speed of light in vacuum. When multiple frequencies are available, the dispersive terms can be fully or partially measured thanks to the frequency dependence of the plasma refractive index<sup>10,22</sup>.

Owing to the difference in the transponding ratios of the X band and the Ka band (880/749 and 3,360/3,599, respectively), the overall plasma contribution in PJ3 and PJ6 can be estimated to 75% accuracy<sup>10</sup>. Under the assumption  $\dot{I}_u = \dot{I}_d$

(which has been verified because the Io torus is only 1.5 light-seconds away from Juno), the frequency shift due to the Io torus is obtained by differencing the relative frequency shift of the X and Ka bands, which is described by equation (1):

$$k \left( \frac{1}{f_K^2} + \frac{1}{\alpha_K^2 f_K^2} \right) \dot{I} = \left( \frac{f_K^2 \alpha_K^2 \alpha_X^2 + 1}{f_X^2 \alpha_X^2 \alpha_K^2 + 1} - 1 \right)^{-1} \left\{ y_X - y_K - k \left[ \dot{P}_u \left( \frac{1}{f_X^2} - \frac{1}{f_K^2} \right) + \dot{P}_d \left( \frac{1}{\alpha_X^2 f_X^2} - \frac{1}{\alpha_K^2 f_K^2} \right) \right] \right\} \quad (2)$$

In equation (2), the estimated Io torus signal is contaminated by the uplink and downlink interplanetary plasma variations in the columnar electron content. In the PJ3 and PJ6 data, we observed a residual plasma noise of about  $8 \times 10^{-7} \text{ m s}^{-1}$  (relative frequency shift  $2.7 \times 10^{-15}$ ) for 60 s integration time. We assessed the effect of this error by means of numerical simulations.

Simulated Doppler observables of PJ3 and PJ6 were generated using the same dynamical model adopted in the analysis of the PJ3 and PJ6 data. A white Gaussian noise with a standard deviation equal to the observed one was added to the simulated observables. Then, we added a signal that mimics the effect of the Io torus to the simulated Doppler observables using a simple Gaussian model for the path delay  $\Delta I$  on a signal of frequency  $f$ :

$$\Delta I = \Delta I_K \left( \frac{f_K}{f} \right)^2 \exp \left[ -\frac{1}{2} \left( \frac{t - \Delta\tau}{\tau/6} \right)^2 \right] \quad (3)$$

Here  $\Delta I_K$  is the maximum path delay on a signal with frequency  $f_K$ ,  $\tau$  is the total duration of the torus signal (corresponding to 6 standard deviations of a Gaussian curve), and  $\Delta\tau$  is the delay between the time of the maximum path delay and the orbit pericentre. The values of the parameters adopted for each flyby were derived from direct measurements carried out in PJ3 and PJ6. PJ3 (PJ6) observations gave the values  $\Delta I_K = 2.1 \text{ cm}$  (4.6 cm),  $\tau = 120 \text{ min}$  (150 min) and  $\Delta\tau = -15 \text{ min}$  (+10 min). The fractional frequency shift  $\Delta y$  on the Doppler observables is given by:

$$\Delta y = \frac{\Delta I}{c} = - \left( \frac{f_X}{f} \right)^2 \frac{\Delta I_K}{c\tau/6} \frac{t - \Delta\tau}{\tau/6} \exp \left[ -\frac{1}{2} \left( \frac{t - \Delta\tau}{\tau/6} \right)^2 \right] \quad (4)$$

To simulate the calibration error due to the residual plasma noise in equation (2), the calibrations were generated using the same model, but by perturbing the input parameters with white, Gaussian random values. The standard deviations of the

perturbing terms were chosen to match the observed solar plasma noise. The resulting standard deviation  $\delta$  of the path delay was less than 10%.

We then carried out a Monte Carlo simulation using 1,000 noise realizations and obtained a sample of estimated gravity fields. None of the gravity harmonic coefficients changed by more than  $1\sigma$  (Extended Data Figs 4 and 5). By contrast, the Io torus can cause biases up to about  $5\sigma$  on gravity solutions based on X-band data. The most affected gravity coefficients are  $J_2$ ,  $J_3$  and  $J_4$ .

**Tesseral gravity field.** The solution reported in Table 1 includes only degree-2 tesseral gravity harmonics. Although higher-degree tesseral harmonics are not required to fit the data to the noise level, a higher-degree field is certainly present. To assess the effect of a tesseral field on the actual estimate, simulations with synthetic Doppler data were conducted. Thermal wind models with a scale height of 1,900 km (which is consistent with the observed odd harmonics<sup>3</sup>) but with a different scale height for the vortices (associated with the tesseral component), were used to generate synthetic gravity fields. The resulting simulated Doppler observables were fitted with the dynamical model used to obtain our solution (Table 1), limited to degree-2 tesseral harmonics. Our goal was to identify the largest tesseral field (and therefore the largest scale height) that can be hidden in the Doppler data without producing signatures in the residuals. We found that the threshold value of the scale height is about 380 km.

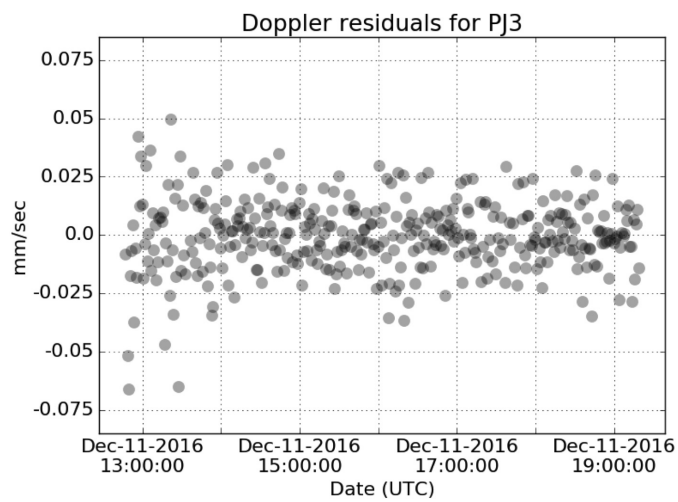
To include the effect of the neglected tesseral field in the estimation, we performed a consider analysis, which quantifies the effect of non-estimated parameters (the higher-degree tesseral field) on the uncertainties of the estimated parameters. The analysis revealed that inclusion of the tesseral field increases these uncertainties. Extended Data Table 2 presents the consider uncertainties of the estimate for a thermal wind model with a vortex scale height of 380 km.

**Data availability.** The Juno tracking data and the ancillary information used in this analysis are archived at NASA's Planetary Data System (<https://pds.nasa.gov>).

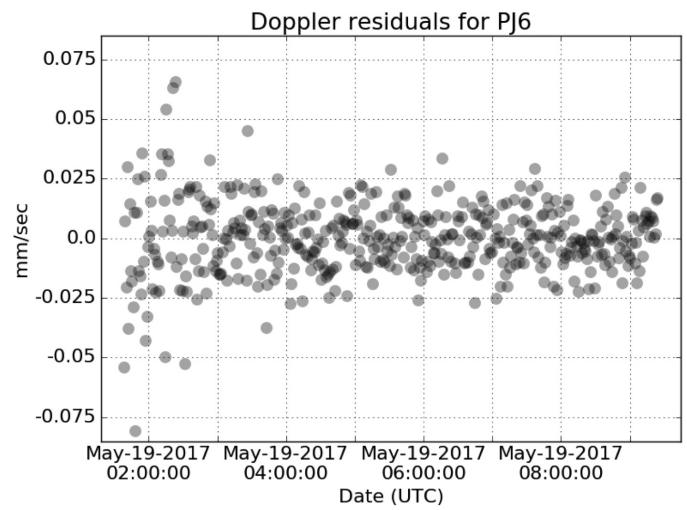
**Code availability.** The analysis presented in this work relies on proprietary orbit determination codes that are not publicly available. The MONTE software package is used at the Jet Propulsion Laboratory for planetary spacecraft navigation. The ORACLE orbit determination filter was developed at Sapienza University of Rome under contract with the Italian Space Agency.

22. Mariotti, G. & Tortora, P. Experimental validation of a dual uplink multifrequency dispersive noise calibration scheme for deep space tracking. *Radio Sci.* **48**, 111–117 (2013).
23. Delamere, P. A. & Bagenal, F. Modeling variability of plasma conditions in the Io torus. *J. Geophys. Res. Space Phys.* **108**, 1276 (2003).
24. Seidelmann, P. K. & Divine, N. Evaluation of Jupiter longitudes in System III (1965). *Geophys. Res. Lett.* **4**, 65–68 (1977).

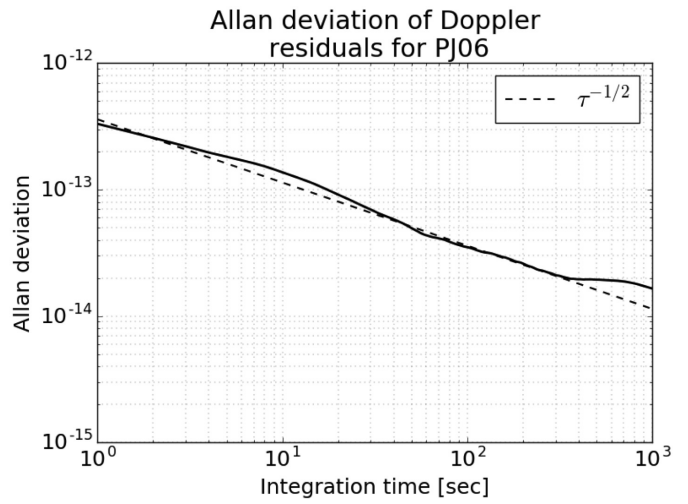
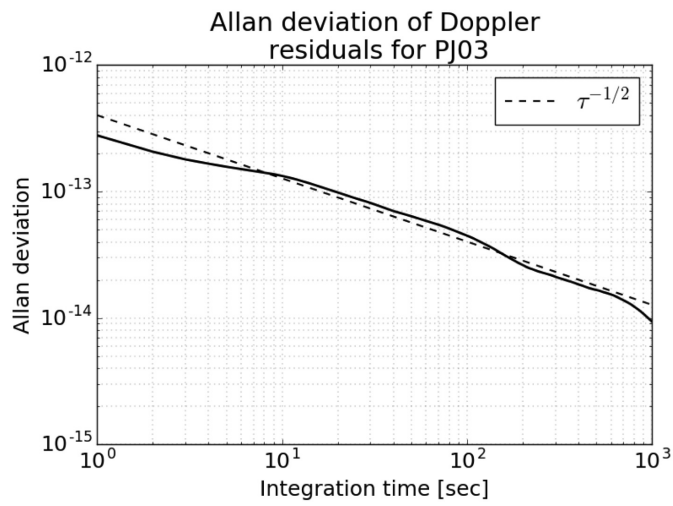




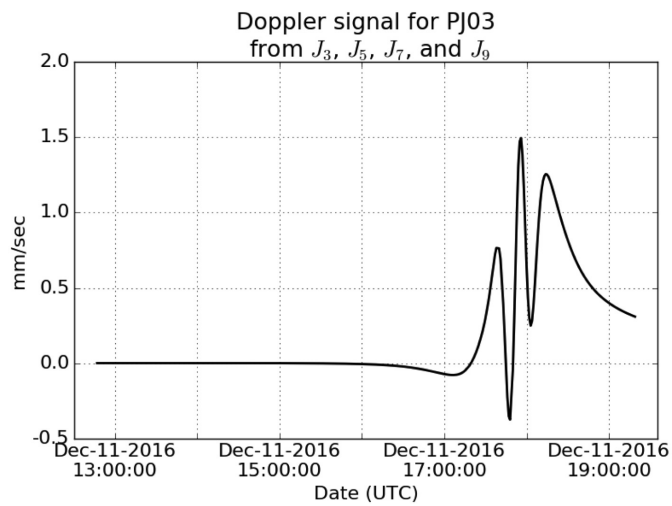
**Extended Data Figure 1 | Range-rate residuals.** Two-way range-rate residuals (integrated over 60 s) for the Ka-band perijove passes PJ3 and PJ6 are shown. The root-mean-square value of the residuals is  $0.015 \text{ mm s}^{-1}$



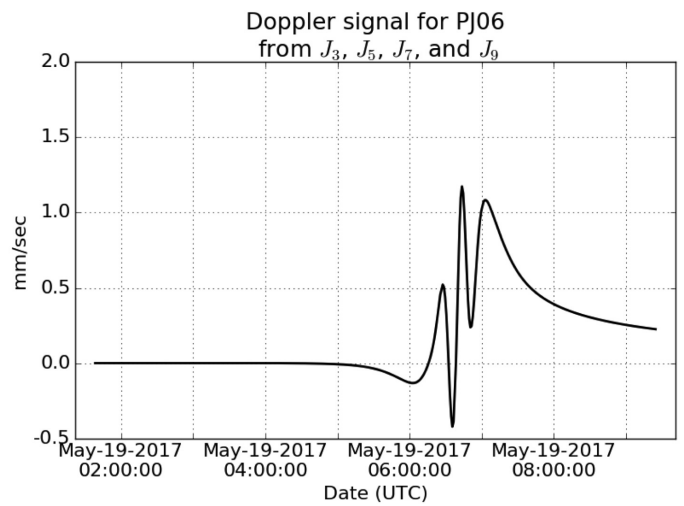
for both passes. The measured range rate was obtained from the radio-science open-loop receiver.



**Extended Data Figure 2 | Frequency stability.** The Allan deviation of relative frequency shift for the Ka-band perijove passes PJ3 and PJ6 is shown. The slopes are roughly consistent with white noise (dashed line).

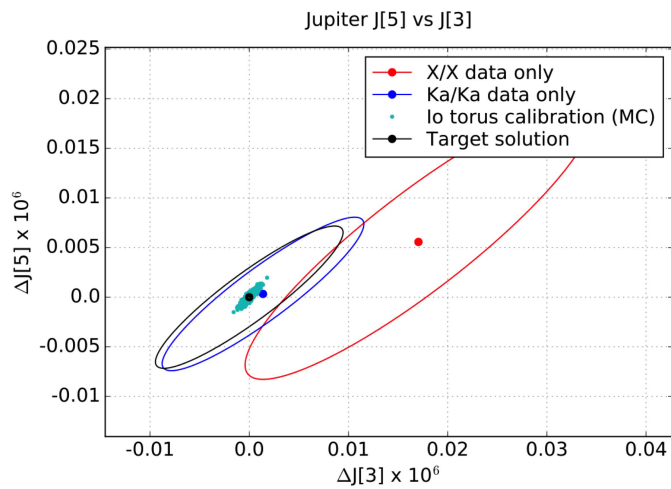


**Extended Data Figure 3 | Gravity harmonic signatures.** Range-rate signals from the  $J_3$ ,  $J_5$ ,  $J_7$  and  $J_9$  gravity harmonics for PJ3 and PJ6 are shown. The smaller signal in PJ6 is due to a less favourable projection of the spacecraft velocity along the Earth–Jupiter line of sight (the angle



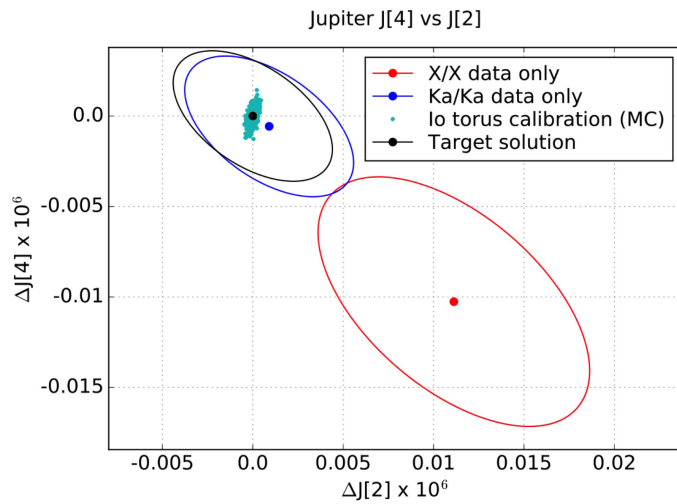
between Juno's orbit normal and the line of sight was  $19.2^\circ$  in PJ3 and  $15.1^\circ$  in PJ6). By comparison, the range-rate noise at 60 s is  $0.015 \text{ mm s}^{-1}$  in both passes.





#### Extended Data Figure 4 | Io torus effects on the estimation of $J_3$ – $J_5$ .

Shown are estimation biases on  $J_3$  and  $J_5$  due to calibration errors of the Io torus path delay variation (cyan dots) in a Monte Carlo (MC) simulation of passes PJ3 and PJ6 of the Juno gravity experiment. The calibration errors are compared to the estimated  $3\sigma$  uncertainty ellipses of the target solution (black), obtained without the Io torus, and the solutions obtained using only X- (red) and Ka-band (blue) data. The estimation bias on  $J_3$  is about  $3\sigma$  if X-band data are used. Ka-band data or dual-link calibration reduce the bias to less than  $1\sigma$ .



#### Extended Data Figure 5 | Io torus effects on the estimation of $J_2$ – $J_4$ .

Shown are estimation biases on  $J_2$  and  $J_4$  from the Monte Carlo simulation, as in Fig. 1. The estimation bias on  $J_2$  and  $J_4$  is larger than  $4\sigma$  if X-band data are used, while using Ka-band or plasma-calibrated data reduces it to less than  $1\sigma$ .

**Extended Data Table 1 | Characteristics of perijove passes PJ3 and PJ6 used in the gravity solution**

	PJ3	PJ6
Perijove epoch	11-DEC-2016 17:03:41 UTC	19-MAY-2017 06:00:44 UTC
One-way light-time	2924 s	2339 s
Perijove latitude	5.7°	8.5°
Perijove altitude	4154 km	3503 km
Sun-Earth-Probe angle	61.6°	135.4°
NON to Earth angle	19.2°	15.1°
Longitude at equator crossing	6.8°	142.0°

Altitude refers to the oblate planet. The negative orbit normal (NON) to Earth is the angle between the opposite of the orbit normal and Earth's direction. Longitude at equator crossing refers to System III<sup>24</sup>.



**Extended Data Table 2 | Consider analysis covariances ( $3\sigma$ )**

	Value	Uncertainty	Consider uncertainty
$J_2$ ( $\times 10^{-6}$ )	14696.572	0.014	0.024
$C_{21}$ ( $\times 10^{-6}$ )	-0.013	0.015	0.018
$S_{21}$ ( $\times 10^{-6}$ )	-0.003	0.026	0.035
$C_{22}$ ( $\times 10^{-6}$ )	0.000	0.008	0.023
$S_{22}$ ( $\times 10^{-6}$ )	0.000	0.011	0.046
$J_3$ ( $\times 10^{-6}$ )	-0.042	0.010	0.024
$J_4$ ( $\times 10^{-6}$ )	-586.609	0.004	0.008
$J_5$ ( $\times 10^{-6}$ )	-0.069	0.008	0.013
$J_6$ ( $\times 10^{-6}$ )	34.198	0.009	0.012
$J_7$ ( $\times 10^{-6}$ )	0.124	0.017	0.024
$J_8$ ( $\times 10^{-6}$ )	-2.426	0.025	0.026
$J_9$ ( $\times 10^{-6}$ )	-0.106	0.044	0.061
$J_{10}$ ( $\times 10^{-6}$ )	0.172	0.069	0.070
$J_{11}$ ( $\times 10^{-6}$ )	0.033	0.112	0.148
$J_{12}$ ( $\times 10^{-6}$ )	0.047	0.178	0.178
$k_{22}$	0.625	0.063	0.118
$\alpha$ ( $^\circ$ )	268.0570	0.0013	0.0052
$\delta$ ( $^\circ$ )	64.4973	0.0014	0.0067

Consider uncertainties are shown after a tesseral field corresponding to a flow depth of 380 km is added to the estimated zonal field in Table 1. Gravity fields generated by larger depths of the tesseral flow would produce signatures in the Doppler residuals<sup>20</sup>.

# Jupiter's atmospheric jet streams extend thousands of kilometres deep

Y. Kaspi<sup>1</sup>, E. Galanti<sup>1</sup>, W. B. Hubbard<sup>2</sup>, D. J. Stevenson<sup>3</sup>, S. J. Bolton<sup>4</sup>, L. Iess<sup>5</sup>, T. Guillot<sup>6</sup>, J. Bloxham<sup>7</sup>, J. E. P. Connerney<sup>8,9</sup>, H. Cao<sup>3,7</sup>, D. Durante<sup>5</sup>, W. M. Folkner<sup>10</sup>, R. Helled<sup>11</sup>, A. P. Ingersoll<sup>3</sup>, S. M. Levin<sup>10</sup>, J. I. Lunine<sup>12</sup>, Y. Miguel<sup>6,13</sup>, B. Militzer<sup>14</sup>, M. Parisi<sup>10</sup> & S. M. Wahl<sup>14</sup>

The depth to which Jupiter's observed east–west jet streams extend has been a long-standing question<sup>1,2</sup>. Resolving this puzzle has been a primary goal for the Juno spacecraft<sup>3,4</sup>, which has been in orbit around the gas giant since July 2016. Juno's gravitational measurements have revealed that Jupiter's gravitational field is north–south asymmetric<sup>5</sup>, which is a signature of the planet's atmospheric and interior flows<sup>6</sup>. Here we report that the measured odd gravitational harmonics  $J_3$ ,  $J_5$ ,  $J_7$  and  $J_9$  indicate that the observed jet streams, as they appear at the cloud level, extend down to depths of thousands of kilometres beneath the cloud level, probably to the region of magnetic dissipation at a depth of about 3,000 kilometres<sup>7,8</sup>. By inverting the measured gravity values into a wind field<sup>9</sup>, we calculate the most likely vertical profile of the deep atmospheric and interior flow, and the latitudinal dependence of its depth. Furthermore, the even gravity harmonics  $J_8$  and  $J_{10}$  resulting from this flow profile also match the measurements, when taking into account the contribution of the interior structure<sup>10</sup>. These results indicate that the mass of the dynamical atmosphere is about one per cent of Jupiter's total mass.

The Juno gravity measurements so far have improved the accuracy of the known gravity harmonics  $J_2$ ,  $J_4$ ,  $J_6$  and  $J_8$  by more than two orders of magnitude<sup>5,11</sup>. These low-degree even gravity harmonics are mostly affected by Jupiter's interior density structure and its shape<sup>12</sup>, and therefore, although the signal from these harmonics may contain a contribution<sup>13</sup> from the atmospheric and interior flows ( $\Delta J_n$ ), it is difficult to use these harmonics to infer information about the flows directly. The gravity measurements also revealed north–south asymmetries in Jupiter's gravity field<sup>5</sup>, which are manifested as large values of the odd gravity harmonics  $J_3$ ,  $J_5$ ,  $J_7$  and  $J_9$  (see Table 1). Because a gas planet rotating as a solid body has no asymmetry between north and south, any non-zero value of the odd  $J_n$  must come from dynamics<sup>6</sup>. As the observed cloud-level flow is not hemispherically symmetric (Fig. 1), if enough mass is involved in the asymmetric component of the flow it will produce large odd  $J_n$ . Although the flow is also expected to dominate the high-degree harmonics<sup>3</sup>, the gravity harmonics beyond  $J_{10}$  are still beneath the level of the measurement uncertainty<sup>5</sup>. In addition, because the low-degree even  $J_n$  are dominated by solid-body rotation, the only current measurements that can be uniquely related to the dynamics are the low-degree odd harmonics  $J_3$  to  $J_9$ . Therefore, in this study, we use only those to infer the depth of the cloud-level winds.

Because Jupiter is rotating with a short period of 9.92 h, the flow within the planet to leading order is in geostrophic balance, meaning that the momentum budget is dominated by the balance between the Coriolis force and the horizontal pressure gradients. As a consequence,

the flow to leading order is in thermal wind balance, namely

$$2\Omega \cdot \nabla(\rho_s \mathbf{u}) = \nabla \rho' \times \mathbf{g} \quad (1)$$

where  $\Omega$  is the planetary rotation rate vector,  $\mathbf{u}$  is the velocity field,  $\rho_s$  and  $\rho'$  are the static and dynamic components of the density, respectively, and  $\mathbf{g}$  is the gravity obtained by integrating  $\rho_s$  (see Methods)<sup>14</sup>. Non-spherical effects can play a part in this balance (for example, the deviation of  $\mathbf{g}$  from radial symmetry)<sup>15,16</sup>; however, it has been shown that to leading order equation (1) captures the dynamical balance well<sup>16,17</sup> (Extended Data Fig. 1). As the gravity harmonics induced by the flow are related to  $\rho'$  directly, we can relate the flow field and the gravity spectrum. Thus, given the measured gravitational field, inversion of equation (1) allows us to infer the flow profile that best matches the measurements. For this inversion we use an optimization based on the adjoint method<sup>9</sup> (see Methods).

The relation between the odd gravity harmonics and the flow is shown in Fig. 2 for a simple model<sup>6</sup> where the depth of the cloud-level wind is parameterized with a single decay parameter,  $H$ . In this scenario, the interior flow is an extension of the cloud-level flow, along the direction of the spin axis owing to angular momentum constraints (see below)<sup>14,18</sup>, but decaying exponentially in radius with  $H$  being the e-folding decay depth<sup>6,19</sup>. The Juno-measured values (Fig. 2, dashed lines), show that for all four harmonics, independently, the theoretical values<sup>6</sup> capture the correct sign of the measured harmonics and indicate that the e-folding decay depth of the flow is between 1,000 km and 3,000 km (Fig. 2, grey shading). Inverting the gravity field<sup>9</sup>, taking into consideration the uncertainties of each of the measured harmonics and their cross-correlated uncertainties (the error covariance matrix, see Methods), gives an e-folding decay depth of about 1,500 km. We note, however, that the measured value of  $J_5$  deviates by a factor of about two from the corresponding theoretical value of a single-parameter deep wind profile, suggesting that a more elaborate vertical flow profile than the simple exponential decay is needed to match the data.

Given that the measurements provide four non-zero odd gravity harmonics, a more complex optimization of the vertical and meridional

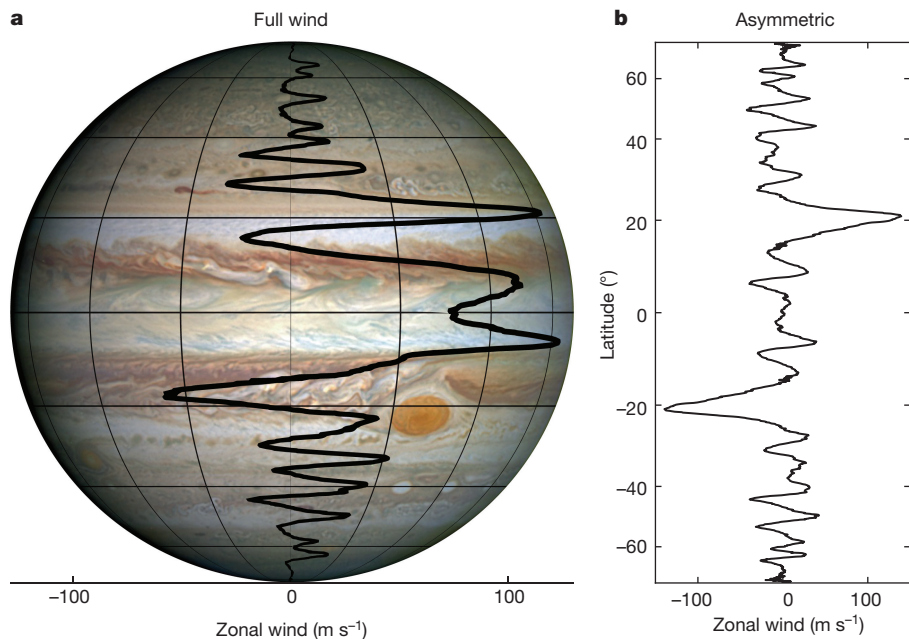
**Table 1 | The Juno-measured and model odd gravity harmonics**

Harmonic	Measured	Model without latitudinal variation	Model with latitudinal variation
$J_3 (\times 10^{-8})$	$-4.24 \pm 0.91$	$-5.71 \pm 1.67$	$-5.96 \pm 2.33$
$J_5 (\times 10^{-8})$	$-6.89 \pm 0.81$	$-7.73 \pm 0.41$	$-8.00 \pm 0.43$
$J_7 (\times 10^{-8})$	$12.39 \pm 1.68$	$12.77 \pm 0.54$	$12.04 \pm 0.70$
$J_9 (\times 10^{-8})$	$-10.58 \pm 4.35$	$-8.84 \pm 0.42$	$-9.71 \pm 0.72$

Model results are shown for optimizations with and without variation of flow depth with latitude. The uncertainties are the  $3\sigma$  uncertainty values. The model uncertainty is calculated by the optimization procedure (Methods). For the middle (right) column the  $J_n$  values correspond to the parameter values given in the caption of Fig. 3 (Fig. 4).

<sup>1</sup>Department of Earth and Planetary Sciences, Weizmann Institute of Science, Rehovot 76100, Israel. <sup>2</sup>Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA.

<sup>3</sup>Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. <sup>4</sup>Southwest Research Institute, San Antonio, Texas 78238, USA. <sup>5</sup>Department of Mechanical and Aerospace Engineering, Sapienza Università di Roma, 00184 Rome, Italy. <sup>6</sup>Université Côte d'Azur, OCA, Lagrange CNRS, 06304 Nice, France. <sup>7</sup>Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA. <sup>8</sup>Space Research Corporation, Annapolis, Maryland 21403, USA. <sup>9</sup>NASA/GSFC, Greenbelt, Maryland 20771, USA. <sup>10</sup>Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. <sup>11</sup>Institute for Computational Science, Center for Theoretical Astrophysics and Cosmology, University of Zurich, 8057 Zurich, Switzerland. <sup>12</sup>Department of Astronomy, Cornell University, Ithaca, New York 14853, USA. <sup>13</sup>Leiden Observatory, University of Leiden, Leiden, The Netherlands. <sup>14</sup>Department of Earth and Planetary Science, University of California, Berkeley, California 94720, USA.



**Figure 1 | Jupiter's asymmetric zonal velocity field.** **a**, The cloud-level zonal flows (thick black line) as a function of longitude, as measured during Juno's third perijove pass on 11 December 2016 (ref. 30). The image of Jupiter was taken by the Hubble Wide Field Camera in 2014 (<https://en.wikipedia.org/wiki/Jupiter>). Grid latitudes are as in **b** and the longitudinal spread is 45°. Zonal flow scale is the same as the longitudinal grid on the sphere. **b**, The asymmetric component of the flow, taken as the difference between the northern and southern hemisphere cloud-level flows.

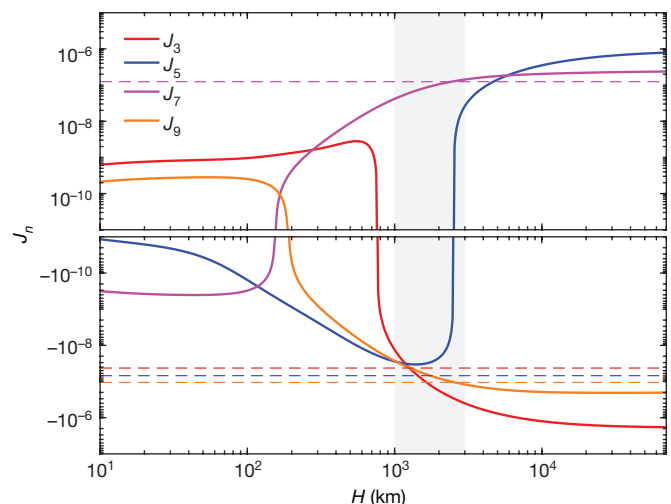
profiles of the zonal flow is indeed feasible. Motivated by the Galileo probe measurement of a relatively constant wind profile<sup>20</sup> between 4 bar and 22 bar, and magnetohydrodynamic theory suggesting that ohmic dissipation will cause a more abrupt decay of the flow at depth<sup>7,8,21,22</sup>, we add—in addition to the exponential decay function used in the first estimate (Fig. 2)—a vertical decay profile expressed as a hyperbolic tangent ('tanh') function and a free parameter  $\alpha$ , representing the ratio between the two functions. This allows for a much wider range of vertical decay profiles, with three free parameters defining the vertical profile of the flow: the depth  $H$  represents the inflection point of the tanh function,  $\Delta H$  represents the decay width of the tanh function and  $\alpha$  is the ratio between the tanh function and an exponential decay with the same decay depth  $H$ . Using these three parameters as control parameters in the inverse adjoint model, the optimization process (Fig. 3) minimizes a cost function, taking into account the uncertainties in the gravity measurements, including the error covariance between the different harmonics (Methods)<sup>9,23</sup>.

Beginning with an assumed vertical decay profile as an initial condition (dashed line in Fig. 3a and black squares in Fig. 3b, c), the optimization iteratively minimizes the cost function, reaching a unique global minimum in the three-dimensional parameter space of  $H$ ,  $\Delta H$  and  $\alpha$  (red dot in Fig. 3b, c). The best optimized solution, defining a particular vertical profile of the zonal flow (red line in Fig. 3a), is achieved with  $H = 1,803 \pm 351$  km,  $\Delta H = 1,570 \pm 422$  km and  $\alpha = 0.92 \pm 0.26$ , where the error is calculated by the optimization process (see Methods), indicating a very deep flow profile containing a large mass. We note that the minimum of the cost function for  $\Delta H$  is rather flat towards lower  $\Delta H$  (Fig. 3b), indicating that a flow profile with a much more abrupt decay at depth is compatible with the measured  $J_n$ . Integrating the density profile  $\rho_s$  down to where the flow decreases noticeably (about 3,000 km) reveals that this region contains about 1% of Jupiter's mass (the mass dependence on depth is shown in Extended Data Fig. 2). This large mass of the dynamical atmosphere (the region that is differentially rotating) is consistent with the persistence of the observed jets over the past several decades<sup>2</sup>. In an accompanying paper<sup>10</sup> we show that, on the basis of the even harmonics, beneath this dynamical atmosphere, in Jupiter's deep interior, there is probably very little zonal flow. The angular momentum of this flow is about  $2 \times 10^{-5}$  that of the solid-body rotating planet.

The solution shown in Fig. 3a (red line) implies that the meridional profile of the flow at depth is strongly correlated with the cloud-level flow. To test the statistical significance of this solution we generate a

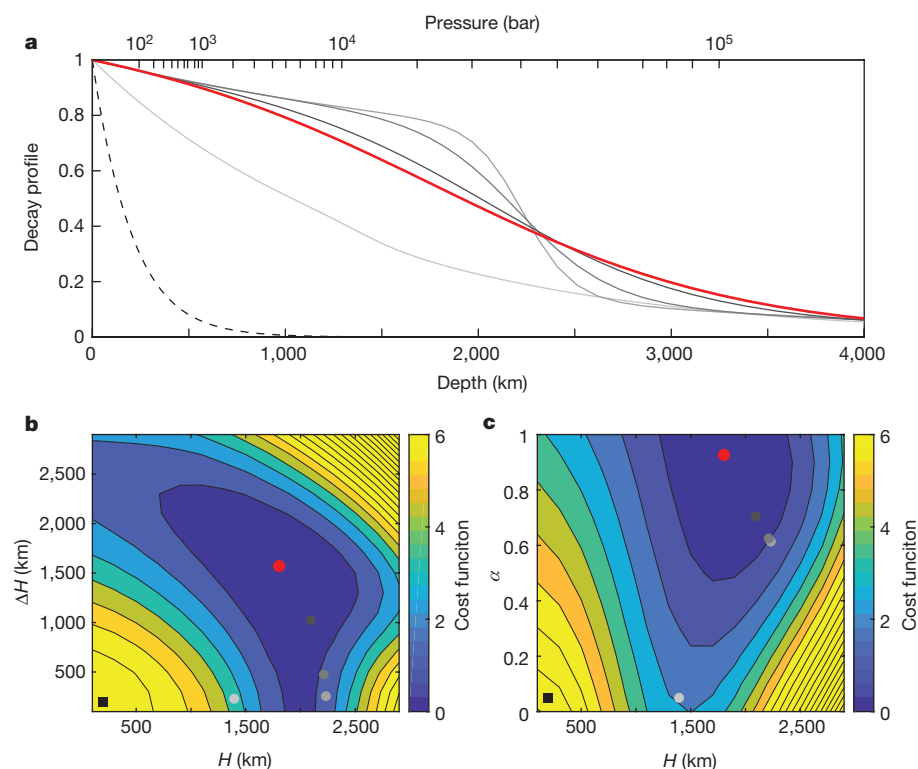
large set of synthetic zonal wind profiles (Extended Data Fig. 3) by expanding the observed flow up to high-degree Legendre polynomials and summing them back up while assigning random signs to the expansion coefficients. We find that the solution using the observed cloud-level wind profile (Extended Data Fig. 4, black) is one of the closest solutions to the measurements (Extended Data Fig. 4, red) and only a very small subset of the random flow profiles (less than 1%) give a lower cost-function value (Extended Data Fig. 4, green). This shows that it is statistically improbable that the meridional profile of the flow changes with depth, or that the solution was found by chance (see further discussion in Methods).

Considering the angular momentum budget is helpful for developing a mechanistic understanding of these deep dynamics. Modelling studies



**Figure 2 | The odd gravity harmonics as function of a single e-folding decay depth parameter  $H$ .** The predicted values<sup>6</sup> (solid) and the Juno-measured values<sup>5</sup> (dashed, corresponding to the values in Table 1) for  $J_3$  (red),  $J_5$  (blue),  $J_7$  (magenta) and  $J_9$  (orange) are shown as functions of  $H$ . All four gravity harmonic measurements, independently, indicate that the e-folding depth of the flow is 1,000–3,000 km (grey shading). All four odd harmonics are small if the flows are shallow, and become large for deeper flows that contain more mass. The change in sign at different decay depths depends on the way the flow pattern projects onto the different Legendre polynomials.





**Figure 3 | Jupiter's optimized vertical profile of the zonal wind.** **a**, The vertical profile of the flow from the optimization process, beginning with an initial profile (dashed), which evolves along the optimization process (from light to dark shades of grey), leading to the best optimized vertical profile (red), with the parameters  $H = 1,803 \pm 351$  km,  $\Delta H = 1,570 \pm 422$  km and  $\alpha = 0.92 \pm 0.26$  (equation (13) in Methods). The abscissa shows both the depth (bottom) and pressure (top) beneath the 1-bar level. **b**, The cost function in the plane of  $H$  and  $\Delta H$  showing a robust minimum at  $H = 1,803$  km and  $\Delta H = 1,570$  km (red dot). **c**, The cost function in the plane of  $H$  and  $\alpha$  showing a minimum at  $H = 1,803$  km and  $\alpha = 0.92$  (red dot). In **b** and **c** the grey-shaded dots correspond to the grey-shaded curves in **a**. Cost-function values in the colour scale are divided by 1,000 (see calculation in Methods). A statistical significance test for the latitudinal dependence of the flow profile appears in Extended Data Figs 4 and 5.

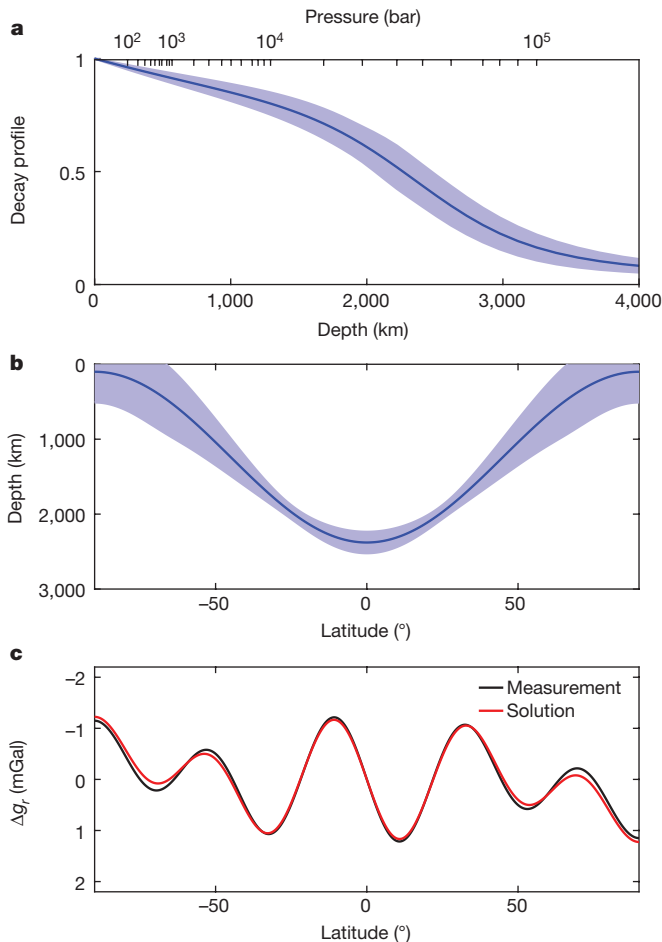
have suggested<sup>18,21</sup> that the leading-order angular momentum balance is  $\mathbf{u} \cdot \nabla M = D - S$ , where  $\mathbf{u}$  is the mass-averaged velocity,  $M$  is the total angular momentum,  $D$  is the drag due to the Lorentz force at depth and  $S = (1/\rho) \nabla \cdot (\rho \mathbf{u}' M')$  is the eddy angular momentum flux divergence, with the overbar indicating a zonal and temporal mean. At the observed cloud level, the eastward (westward) jets are correlated with regions of eddy momentum flux convergence (divergence), that is, where  $S$  is negative (positive)<sup>21,24</sup>. Below that, where the eddy momentum flux convergence is expected to become weak<sup>24</sup>, that is,  $\mathbf{u} \cdot \nabla M \approx 0$ , the flow is along angular momentum surfaces, which on Jupiter are almost entirely parallel to the axis of rotation<sup>14,18,25</sup>. Then, in the deep region, where the fluid becomes electrically conducting (mainly due to pressure ionization) and the Lorentz force may become important (depending on the magnetic field structure) the leading-order balance is  $\mathbf{u} \cdot \nabla M = D$  and the circulation closes. Kinematic dynamo models, which calculate the magnetic drag at depth on the basis of the radially varying electric conductivity inside Jupiter, find that the depth at which the Lorentz drag  $D$  becomes important<sup>7,8</sup> is about 3,000 km. Thus, the theoretical magnetic field considerations and the gravity measurements, which are completely independent, give very consistent results.

Three-dimensional hydrodynamic models of Jupiter, driven by shallow atmospheric turbulence<sup>21,26</sup> or deep internal convection<sup>14</sup>, have found that the low latitudes are often more barotropic than the high latitudes. Thus, an additional level of complexity that can be added to the optimization is allowing the decay depth  $H$  to vary with latitude. To limit the number of optimized parameters, the decay depth is expanded in Legendre polynomials to second order, increasing the number of optimized parameters to four (see Methods). Similarly to the case of a latitudinally independent vertical profile (Fig. 3), in this case the optimized vertical decay profile is rather barotropic at lower depths and extends to great depth (Fig. 4a). The optimization uncertainty is shown graphically by the blue shading, with the values for the profile at the equator given in the caption. At higher latitudes, the vertical decay occurs at shallower depths, and the associated uncertainty grows to approximately 500 km (Fig. 4b). The values of  $J_n$  corresponding to the solutions of Figs 3 and 4 appear in Table 1. We note that with more free parameters than used in these optimizations, closer matches

to the measurements can be reached. However, the power of these solutions is that they are based on relatively simple extensions of the cloud-level flow, giving results remarkably close to all four independent gravity measurements; and, regardless of the exact vertical profile, the solutions indicate that the observed cloud-level flows extend to depths of thousands of kilometres.

The flow profile determined by the odd harmonics also has a signature in the even harmonics. Owing to the uncertainty in the bulk interior density structure of Jupiter<sup>10,27</sup>, there is a wide range of solutions for the low-degree static gravity harmonics  $J_n^s$ , which does not allow us to test uniquely whether the  $\Delta J_n$  from the even harmonics matches the measured values via  $\Delta J_n = J_n - J_n^s$ . However, for  $J_8$  and  $J_{10}$  the interior models are very constraining<sup>10</sup>, giving values between  $-245.7 \times 10^{-8}$  and  $-246.3 \times 10^{-8}$  for  $J_8^s$ , and between  $20.1 \times 10^{-8}$  and  $20.4 \times 10^{-8}$  for  $J_{10}^s$  (for interior models that also match  $J_4$  and  $J_6$ ). The measured Juno values are  $J_8 = (-242.6 \pm 0.8) \times 10^{-8}$  and  $J_{10} = (17.2 \pm 2.3) \times 10^{-8}$ , meaning that a positive (negative) correction by the dynamics is needed to match the measurements for  $J_8$  ( $J_{10}$ ). The values corresponding to the flow profiles presented in Figs 3 and 4 (Extended Data Table 1) are indeed such that for both cases, and for both  $J_8$  and  $J_{10}$ , the dynamical corrections can reconcile the differences between the measurements and the internal models, further confirming that the inferred flow profile presented here matches the measurements from Juno. An accompanying paper<sup>10</sup> shows that using the range of current interior models gives further constraints on possible deeper interior flow.

Juno's gravity measurements are consistent with Juno's microwave radiometer measurements, indicating a north–south asymmetry in the sub-cloud-level atmospheric composition, and a direct signature of the main equatorial belt to the maximum depth of the microwave sensitivity<sup>11,28</sup> at about 1,000 bar. With more Juno orbits the microwave measurements<sup>4,29</sup> will obtain greater and improved thermal mapping of the deep atmosphere, which will better constrain the water and ammonia abundances as well as the atmospheric flows at those levels. As the Juno mission completes its global mapping of Jupiter, the combination of the gravity, magnetic and microwave data may provide further insights into the coupling between Jupiter's deep interior and atmospheric flows.



**Figure 4 | Jupiter's optimized vertical profile of the zonal wind when allowing for its latitudinal variation.** **a**, The vertical profile of the flow at the equator from the optimization process (blue line) and its uncertainty (blue shading). The best optimized values at the equator are  $H = 2,379 \pm 142$  km,  $\Delta H = 819 \pm 437$  km and  $\alpha = 0.62 \pm 0.09$ . The abscissa shows both the depth (bottom) and pressure (top) beneath the 1-bar level. **b**, The variation of the inflection point (as shown in **a**) with latitude (blue line) and its uncertainty (blue shading). Details of the latitudinal dependence of  $H$  and its functional form are given in Methods (equation (13)). **c**, The Juno measurement of the asymmetric gravity field  $\Delta g_l$  (for  $J_3$ – $J_9$ ) as a function of latitude and the corresponding values from the best-fit solution (**a** and **b**), showing a good match between the measurements and the optimized solution (see calculation in Methods).

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 September 2017; accepted 17 January 2018.

1. Dowling, T. E. Dynamics of Jovian atmospheres. *Annu. Rev. Fluid Mech.* **27**, 293–334 (1995).
2. Vasavada, A. R. & Showman, A. P. Jovian atmospheric dynamics: an update after Galileo and Cassini. *Rep. Prog. Phys.* **68**, 1935–1996 (2005).
3. Hubbard, W. B. Gravitational signature of Jupiter's deep zonal flows. *Icarus* **137**, 357–359 (1999).
4. Bolton, S. J. *Juno Final Concept Study Report*. Technical Report AO-03-OSS-03 (New Frontiers, NASA, 2005).
5. Less, L. *et al.* Measurement of Jupiter's asymmetric gravity field. *Nature* **555**, <https://doi.org/10.1038/nature25776> (2018).
6. Kaspi, Y. Inferring the depth of the zonal jets on Jupiter and Saturn from odd gravity harmonics. *Geophys. Res. Lett.* **40**, 676–680 (2013).
7. Liu, J., Goldreich, P. M. & Stevenson, D. J. Constraints on deep-seated zonal winds inside Jupiter and Saturn. *Icarus* **196**, 653–664 (2008).
8. Cao, H. & Stevenson, D. J. Zonal flow magnetic field interaction in the semi-conducting region of giant planets. *Icarus* **296**, 59–72 (2017).
9. Galanti, E. & Kaspi, Y. An adjoint based method for the inversion of the Juno and Cassini gravity measurements into wind fields. *Astrophys. J.* **820**, 91 (2016).

10. Guillot, T. *et al.* A suppression of differential rotation in Jupiter's deep interior. *Nature* **555**, <https://doi.org/10.1038/nature25775> (2018).
11. Bolton, S. J. *et al.* Jupiter's interior and deep atmosphere: the initial pole-to-pole passes with the Juno spacecraft. *Science* **356**, 821–825 (2017).
12. Hubbard, W. B. High-precision Maclaurin-based models of rotating liquid planets. *Astrophys. J.* **756**, L15 (2012).
13. Kaspi, Y. *et al.* The effect of differential rotation on Jupiter's low-degree even gravity moments. *Geophys. Res. Lett.* **44**, 5960–5968 (2017).
14. Kaspi, Y., Flierl, G. R. & Showman, A. P. The deep wind structure of the giant planets: results from an anelastic general circulation model. *Icarus* **202**, 525–542 (2009).
15. Zhang, K., Kong, D. & Schubert, G. Thermal-gravitational wind equation for the wind-induced gravitational signature of giant gaseous planets: mathematical derivation, numerical method and illustrative solutions. *Astrophys. J.* **806**, 270–279 (2015).
16. Cao, H. & Stevenson, D. J. Gravity and zonal flows of giant planets: from the Euler equation to the thermal wind equation. *J. Geophys. Res. Planets* **122**, 686–700 (2017).
17. Galanti, E., Kaspi, Y. & Tziperman, E. A full, self-consistent, treatment of thermal wind balance on fluid planets. *J. Comput. Phys.* **310**, 175–195 (2017).
18. Schneider, T. & Liu, J. Formation of jets and equatorial superrotation on Jupiter. *J. Atmos. Sci.* **66**, 579–601 (2009).
19. Kaspi, Y., Hubbard, W. B., Showman, A. P. & Flierl, G. R. Gravitational signature of Jupiter's internal dynamics. *Geophys. Res. Lett.* **37**, L01204 (2010).
20. Atkinson, D. H., Pollack, J. B. & Seiff, A. The Galileo probe Doppler wind experiment: measurement of the deep zonal winds on Jupiter. *J. Geophys. Res.* **103**, 22911–22928 (1998).
21. Liu, J. & Schneider, T. Mechanisms of jet formation on the giant planets. *J. Atmos. Sci.* **67**, 3652–3672 (2010).
22. Liu, J., Schneider, T. & Kaspi, Y. Predictions of thermal and gravitational signals of Jupiter's deep zonal winds. *Icarus* **224**, 114–125 (2013).
23. Galanti, E. & Kaspi, Y. Deciphering Jupiter's deep flow dynamics using the upcoming Juno gravity measurements and a dynamical inverse model. *Icarus* **286**, 46–55 (2017).
24. Salyk, C., Ingersoll, A. P., Lorre, J., Vasavada, A. & Del Genio, A. D. Interaction between eddies and mean flow in Jupiter's atmosphere: analysis of Cassini imaging data. *Icarus* **185**, 430–442 (2006).
25. Busse, F. H. A simple model of convection in the Jovian atmosphere. *Icarus* **29**, 255–260 (1976).
26. Lian, Y. & Showman, A. P. Generation of equatorial jets by large-scale latent heating on the giant planets. *Icarus* **207**, 373–393 (2010).
27. Wahl, S. *et al.* Comparing Jupiter interior structure models to Juno gravity measurements and the role of an expanded core. *Geophys. Res. Lett.* **44**, 4649–4659 (2017).
28. Li, C. *et al.* The distribution of ammonia on Jupiter from a preliminary inversion of Juno microwave radiometer data. *Geophys. Res. Lett.* **44**, 5317–5325 (2017).
29. Janssen, M. A. *et al.* Microwave remote sensing of Jupiter's atmosphere from an orbiting spacecraft. *Icarus* **173**, 447–453 (2005).
30. Tollefson, J. *et al.* Changes in Jupiter's zonal wind profile preceding and during the Juno mission. *Icarus* **296**, 163–178 (2017).

**Acknowledgements** We thank M. Allison and A. Showman for discussions. The research described here was carried out in part at the Weizmann Institute of Science (WIS) under the sponsorship of the Israeli Space Agency, the Helen Kimmel Center for Planetary Science at the WIS and the WIS Center for Scientific Excellence (Y.K. and E.G.); at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA (W.M.F., M.P. and S.M.L.); at the Southwest Research Institute under contract with NASA (S.J.B.); at the Université Côte d'Azur under the sponsorship of Centre National d'Études Spatiales (T.G. and Y.M.); and at La Sapienza University under contract with Agenzia Spaziale Italiana (L.I. and D.D.). All authors acknowledge support from the Juno project.

**Author Contributions** Y.K. and E.G. designed the study. Y.K. wrote the paper. E.G. developed the gravity inversion model. D.J.S. led the working group within the Juno Science Team and provided theoretical support. W.B.H. initiated the Juno gravity experiment and provided theoretical support. W.B.H., T.G., Y.M., R.H., B.M. and S.L.W. provided interior models and tested the implications of the results. L.I., D.D., W.M.F. and M.P. carried out the analysis of the Juno gravity data. H.C., D.J.S. and J.B. supported the interpretation regarding the magnetic field. J.I.L. and A.P.I. provided theoretical support. S.J.B., S.M.L. and J.E.P.C. supervised the planning, execution and definition of the Juno gravity experiment. All authors contributed to the discussion and interpretation of the results within the Juno Interiors Working Group.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Y.K. (yohai.kaspi@weizmann.ac.il).

**Reviewer Information** *Nature* thanks J. Fortney and N. Nettelmann for their contribution to the peer review of this work.

## METHODS

**Calculation of the dynamical gravity harmonics.** The gravity harmonics  $J_n$  are defined as a weighted integral over the interior density distribution  $J_n = -(Ma^n)^{-1} \int P_n \rho r^n d^3r$ , where  $M$  is the planetary mass,  $a$  is the equatorial radius,  $P_n$  is the  $n$ th Legendre polynomial,  $\rho$  is the local density and  $r$  is the local radius<sup>31</sup>. On planets with internal dynamics, the density is perturbed by the flow so that the total density in  $J_n$  can be written as  $\rho = \rho_s + \rho'$ , where the density  $\rho_s$  is the hydrostatic density resulting from the background rotation and internal density distribution<sup>27,32–35</sup>, and  $\rho'$  are the density fluctuations arising from the atmospheric and internal dynamics<sup>19</sup>. The gravity harmonics can be similarly decomposed into two parts  $J_n = J_n^s + \Delta J_n$ , where the static component  $J_n^s$  is due to the planet's internal density distribution and shape<sup>12,36</sup>, and the dynamical component  $\Delta J_n$  is due to the density deviations related to the flow<sup>19</sup>.

To develop the relation between the flow on Jupiter and the gravity field measured by Juno, we consider the full momentum balance on a rotating planet

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} + 2\boldsymbol{\Omega} \times \mathbf{u} + \boldsymbol{\Omega} \times \boldsymbol{\Omega} \times \mathbf{r} = -\frac{1}{\rho} \nabla p + \nabla \Phi \quad (2)$$

where  $\mathbf{u}$  is the three-dimensional flow vector,  $\boldsymbol{\Omega}$  is the planetary rotation rate vector (magnitude,  $1.76 \times 10^{-4} \text{ s}^{-1}$ ),  $\rho$  is the density,  $p$  is the pressure and  $\Phi$  is the body force potential set by gravity<sup>37</sup> so that  $\nabla \Phi = -\mathbf{g}$ . The first term on the left-hand side is the local acceleration of the flow, the second is the Eulerian advection, the third is the Coriolis acceleration and the fourth is the centrifugal acceleration. On the right-hand side are the pressure gradient and the body force. Frictional forces are neglected. For Jupiter parameters and large-scale motion, the Rossby number is small,  $\text{Ro} \equiv U/\Omega L \approx 0.05$ , where  $U$  is the typical value of the velocity ( $\mathcal{O}(100 \text{ m s}^{-1})$ ) and  $L$  is the typical jet scale ( $\mathcal{O}(10^4 \text{ km})$ ). The small Rossby number implies that the first two terms are negligible compared to the Coriolis term, so that

$$2\boldsymbol{\Omega} \times (\rho \mathbf{u}) = -\nabla p - \rho \mathbf{g} - \rho \boldsymbol{\Omega} \times \boldsymbol{\Omega} \times \mathbf{r} \quad (3)$$

Because for Jupiter parameters the ratio between the two latter terms on the right-hand side of equation (3) is  $a\Omega^2/g \approx 0.1$ , and not two orders of magnitude smaller, as it is for Earth parameters, we do not a priori make the traditional approximation merging the centrifugal force with the gravity term<sup>38</sup>, but solve for the full system, allowing the density, pressure and gravity to be functions of radius  $r$  and latitude  $\theta$ . We separate the solution into a static solution in which  $\mathbf{u} = 0$ , with the solutions  $\rho_s(r, \theta)$ ,  $p_s(r, \theta)$  and  $\mathbf{g}_s(r, \theta)$  of the leading-order equation

$$0 = -\nabla p_s - \rho_s \mathbf{g}_s - \rho_s \boldsymbol{\Omega} \times \boldsymbol{\Omega} \times \mathbf{r} \quad (4)$$

and the deviations  $\rho'(r, \theta)$ ,  $p'(r, \theta)$  and  $\mathbf{g}'(r, \theta)$  due to the dynamics, where  $\rho = \rho_s + \rho'$ ,  $p = p_s + p'$  and  $\mathbf{g} = \mathbf{g}_s + \mathbf{g}'$ . For the static part of the solution we use solutions from interior models<sup>27,39</sup>. Subtracting equation (4) from equation (3) gives the leading-order dynamical equation

$$2\boldsymbol{\Omega} \times (\rho_s \mathbf{u}) = -\nabla p' - \rho_s \mathbf{g}' - \rho' \mathbf{g}_s - \rho' \boldsymbol{\Omega} \times \boldsymbol{\Omega} \times \mathbf{r} \quad (5)$$

Taking the curl of equation (5), eliminating the dependence on pressure, yields a single equation in the azimuthal direction

$$\begin{aligned} -2\Omega r \partial_z(\rho_s u) = & -r g_s^{(\theta)} \frac{\partial \rho'}{\partial r} - g_s^{(r)} \frac{\partial \rho'}{\partial \theta} + r \frac{\partial \rho_s}{\partial r} g^{(\theta)(\theta)} - g^{(r)(r)} \frac{\partial \rho_s}{\partial \theta} \\ & - \Omega^2 r \left[ \frac{\partial \rho'}{\partial \theta} \cos^2 \theta + \frac{\partial \rho'}{\partial r} r \cos \theta \sin \theta \right] \end{aligned} \quad (6)$$

where  $u$  is the velocity component in the azimuthal direction, the superscripts  $(r)$  and  $(\theta)$  denote the radial and latitudinal components, respectively, and the notation  $\partial_z \equiv \cos \theta \frac{\partial}{\partial r} + \sin \theta \frac{\partial}{\partial \theta}$  denotes the derivative along the direction of the axis of rotation. Note that this is an integro-differential equation because the gravity  $\mathbf{g}'$  is calculated by integrating  $\rho'$ . Although this equation can be solved numerically<sup>17</sup>, it is very difficult to solve at the required resolution and the approximation below is sufficient for relating the flow field and the gravity harmonics<sup>17</sup>.

A typical solution to equation (6), corresponding to the flow field in Fig. 3, is given in Extended Data Fig. 1. It shows that the leading-order balance is between the left-hand-side term and the second term on the right-hand side of equation (6). All other terms are at least an order of magnitude smaller, and have a very small contribution to the gravitational harmonics<sup>17</sup>. Thus, by taking  $\mathbf{g} = \mathbf{g}_s(r)$  in equation (3) and neglecting the centrifugal term gives the leading-order solution. The curl of equation (3) then gives the leading-order equation—equation (1)—which is a generalized form of the thermal wind equation<sup>14,19</sup>. We note that if a higher correction is desired, all terms in equation (6) must be maintained because the smaller terms in equation (6) partially cancel each other (Extended Data Fig. 1). Approximations not maintaining all these terms would be invalid<sup>15</sup>.

The zonal component of equation (1) is then

$$2\Omega r \partial_z(\rho_s u) = g_s^{(r)} \frac{\partial \rho'}{\partial \theta} \quad (7)$$

which can be integrated to find a solution for the dynamical part of the density given by

$$\rho'(r, \theta) = \frac{2\Omega r}{g_s} \int_{-\pi/2}^{\pi/2} \partial_z(\rho_s(r) u(r, \theta')) d\theta' + \rho'_0(r) \quad (8)$$

where  $\rho'_0(r)$  is an unknown integration function that depends only on radius. Although the density  $\rho'$  cannot be determined uniquely owing to the unknown  $\rho'_0(r)$ , the gravity harmonics due to dynamics

$$\Delta J_n = -\frac{2\pi}{Ma^n} \int_{-\pi/2}^{\pi/2} \cos \theta d\theta \int_0^a r^{n+2} P_n(\sin \theta) \rho'(r, \theta) dr \quad (9)$$

can be determined uniquely since

$$\int_{-\pi/2}^{\pi/2} \cos \theta d\theta \int_0^a r^{n+2} P_n(\sin \theta) \rho'_0(r) dr = 0 \quad (10)$$

To avoid integrating over discontinuities at the equator the integration is performed from the equator poleward in both hemispheres separately<sup>40</sup>. Therefore, given any flow profile, the anomalous density gradient can be determined to leading order (equation (8)) and the resulting dynamical gravity harmonics can be calculated (equation (9)). We note that the sphericity assumption leaves the choice of using the equatorial radius or the mean radius for  $a$ . For consistency with the standard normalization<sup>5,41</sup> of  $J_n$  we use the equatorial radius, but repeating the calculation with the mean radius gives results within one per cent of those presented here.

**Calculation of the gravity anomaly.** Equivalent to the gravity harmonics is the physical gravity anomaly (Fig. 4c), which emphasizes the nature of the solution as function of latitude<sup>19</sup>. The gravity anomaly in the radial direction on the surface of a planet that results from the asymmetric flow is given by

$$\Delta g_r(\theta) = -\frac{GM}{a^2} \sum_n (n+1) \Delta J_n P_n(\sin \theta) \quad (11)$$

where  $G$  is the gravitational constant and  $n = 3, 5, 7$  and  $9$ . In Fig. 4c we show a comparison between the measured<sup>5</sup> and the calculated gravity anomalies. The better match at low latitudes is a result of the measurements having smaller uncertainties at low latitudes owing to the trajectory of the spacecraft, which is at periaapses near Jupiter's lower latitudes during the initial phase of the Juno mission<sup>11,41</sup>.

**Setup of the flow structure.** Our knowledge of the flow field of Jupiter so far comes almost completely from cloud tracking<sup>30,42</sup>. We use this flow field as an upper boundary, and extend the flow into the interior by optimizing the general functions below. Angular momentum constraints require that the flow into the interior follows angular momentum surfaces<sup>14,18,25</sup> (see main text), which on Jupiter are nearly parallel to the direction of the axis of rotation. Magnetic drag<sup>7</sup> and the compressibility of the fluid<sup>14</sup> require that the flow decays at some depth, and therefore we use a flow field with the following general structure

$$u(r, \theta) = u_{\text{cyl}}(s) Q(r) \quad (12)$$

where  $u_{\text{cyl}}(s)$  is the cloud-level azimuthal wind projected downward along the direction of the axis of rotation, and  $s = r \cos(\theta)$  is the distance from the axis of rotation.  $Q(r)$  is the radial decay function we optimize, given by

$$Q(r) = (1 - \alpha) \exp\left(\frac{r - a}{H(\theta)}\right) + \alpha \left[ \frac{\tanh\left(\frac{-a - H(\theta) - r}{\Delta H}\right) + 1}{\tanh\left(\frac{H(\theta)}{\Delta H}\right) + 1} \right] \quad (13)$$

where  $a$  is the planetary radius,  $\alpha$  is the contribution ratio between an exponential and a normalized hyperbolic tangent function and  $\Delta H$  is the width of the hyperbolic tangent. We take a hierarchical approach using this profile at several levels of complexity. First, setting  $\alpha = 0$ , the flow is parameterized as a simple exponential decay, with  $H$  being independent of latitude, as has been done in many previous studies<sup>6,10,19,43,44</sup>. Then, allowing  $0 < \alpha < 1$ , the flow is parameterized (Fig. 3), with three free parameters— $\alpha$ ,  $H$  and  $\Delta H$ —as they appear in equation (13), but still keeping  $H$  as a single number. As a final step (Fig. 4),  $H$  is allowed to vary as a function of latitude and defined as



$$H(\theta) = H_0 + H_2 P_2(\sin\theta) \quad (14)$$

where  $H_0$  is the single latitude-independent depth used in the first and second setups, and  $H_2$  is the additional parameter used to set the amplitude of the latitude-dependent second Legendre polynomial function  $P_2$ . For the optimization shown in Fig. 4 the values are  $H_0 = 1,619 \pm 150$  km and  $H_2 = -1,519 \pm 459$  km. We note that the hyperbolic function is normalized by its value at the surface of the planet to ensure that the surface flow has the value of the measured cloud-level wind. Expansion of  $H(\theta)$  to higher harmonics is possible, but additional optimized parameters increase the solution uncertainty (see below), and therefore we restrict this expansion only to second order.

**The optimization procedure.** The methodology described here is similar to that used in ref. 23. We find the values of a set of control variables that makes the model solution for the gravity harmonics as close as possible to the measured gravity harmonics. The number of optimized control variables in the three setups varies between one and four parameters, as discussed above. The measure for the desired proximity of the model solution to the measurements (a cost function) takes into account our knowledge regarding the observational errors. The optimization procedure provides an efficient way to reach the global minimum of the cost function.

Since  $\alpha$  has different units from those of  $H$  and  $\Delta H$ , the problem is best conditioned when the total control vector is composed from the different parameters normalized by their typical values. We define the general control vector as

$$\mathbf{X}_C = (H_0/h_{\text{nor}}, \Delta H/h_{\text{nor}}, \alpha/\alpha_{\text{nor}}, H_2/h_{\text{nor}})^T \quad (15)$$

where  $h_{\text{nor}} = 10^7$  m and  $\alpha_{\text{nor}} = 1$ . In the optimization procedure, the values of the normalized control variables  $H_0/h_{\text{nor}}$ ,  $\alpha/\alpha_{\text{nor}}$  and  $\Delta H/h_{\text{nor}}$  are limited to the range of 0 to 1, and the value of  $H_2/h_{\text{nor}}$  between  $-1$  and  $1$ .

The cost function is defined as the weighted difference between the model-calculated odd harmonics and those measured by Juno. Together with an additional penalty term to ensure that the initial guess does not affect the solution, the cost function is

$$L = (\mathbf{J}^m - \mathbf{J}^o)^T \mathbf{W} (\mathbf{J}^m - \mathbf{J}^o) + \varepsilon \mathbf{X}_C^T \mathbf{X}_C \quad (16)$$

where  $\mathbf{J}^m = (J_3^m, J_5^m, J_7^m, J_9^m)^T$  is the calculated model solution,  $\mathbf{J}^o = (J_3^o, J_5^o, J_7^o, J_9^o)^T$  is the measured one, and  $\mathbf{W}$  is the  $4 \times 4$  weight matrix (Extended Data Table 2), calculated as the inverse of the covariance matrix multiplied by 9 (equivalent to three times the uncertainties). The diagonal terms give the weight assigned to each harmonic independently, and the off-diagonal terms give the weights resulting for the cross-correlation of the measurement errors. The larger the value, the more weight is given in the cost function. For example, looking at the diagonal terms, the largest weight is given to  $J_5$  and the smallest one to  $J_9$ . Importantly, the off-diagonal terms have values that are as large as the diagonal terms, that is, there is a strong correlation between the measurement errors, and therefore we can expect the discrepancy between the model harmonics and the measured ones also to be cross-correlated in the same manner. The second term in equation (16) acts as a penalty term (also known as 'regularization') whose purpose in this case is to ensure that the optimized solution is not affected by the initial guess, or any part of the control vector that does not affect the difference between the calculated and observed gravity harmonics. An extensive discussion of this issue (also known as the null space of the solution) can be found in previous studies<sup>17,23</sup>. The value of the parameter  $\varepsilon$  is set according to the initial value of the cost function, so it affects the solution only when the cost function is considerably reduced. The form of the penalty term is set to penalize any non-zero value of the control variable  $\mathbf{X}_C$  since there is no prior knowledge of the depth of the flow. Given an initial guess for  $\mathbf{X}_C$ , a minimal value of  $L$  is searched for using the Matlab function 'fmincon' and taking advantage of the cost-function gradient that is calculated with the adjoint of the dynamical model<sup>9</sup>.

**Calculating the uncertainties in the solution.** The control variable uncertainties are derived from the Hessian matrix  $G$  (second derivative of the cost function  $L$  with respect to the control vector  $\mathbf{X}_C$ )<sup>9</sup>. For example, in the third setup of the optimization there are 4 parameters that are optimized, therefore the size of the Hessian matrix will be  $4 \times 4$ . Inverting the Hessian matrix  $G$ , we get the solution error covariance matrix  $C$ . This matrix includes the error covariance associated with combination of each two control variables (off-diagonal terms), and the variance of each one (diagonal terms). Physically, the covariance matrix indicates the formal uncertainties in the control variables given the uncertainties of the observations (weights  $\mathbf{W}$  in the cost function). The larger the uncertainties in the observations, the smaller are the weights in the cost-function, and the larger the uncertainties in the control variables. The uncertainties appearing in this study for  $H$ ,  $\Delta H$  and  $\alpha$  are the square roots of the diagonal terms in the matrix  $C$ . We note that in all cases analysed in this work, the off-diagonal terms in  $C$  have the

same order of magnitude as the diagonal terms, meaning that uncertainties in the control variable are highly correlated.

Using the uncertainties in the control variable, we can calculate the uncertainties in the model solution for  $J_n$ . Since the uncertainties for  $H$ ,  $\Delta H$ , and  $\alpha$  represent the first standard deviation of the errors, we can statistically estimate the associated error in the  $J_n$  values by solving the model with the parameters randomly perturbed around their optimized value (with the perturbations having a normal distribution with the calculated standard deviation). In this study we generate 1,000 such cases, calculate the  $J_n$  for each case, and then calculate the standard deviation for each  $J_n$ . This is the error value given to each gravity harmonic in Table 1 and Extended Data Table 1.

**Statistical significance test for the latitudinal profile.** One of the conclusions of this study is that the observed cloud-level meridional profile of the zonal wind, as observed at the cloud-level, extends deep into the interior. This is a strong constraint on the flow, and we investigate its statistical significance here. Since we are optimizing a solution with only four measurements, there exists a possibility that the match obtained with the gravity measurements is by chance and not because the same meridional profile extends to great depths. To exclude this possibility, we examine whether a match with the gravity measurements could be obtained when using a meridional profile different from that of the cloud-level flows. To make a sensible test, the artificial wind profile we examine should have similar characteristics, such as the typical latitudinal width of the jets and their amplitude. To accomplish this, the observed cloud-level wind is decomposed into the first 100 Legendre polynomials

$$U_{\text{surf}}(\theta) = \sum_{i=0}^{99} A_i P_i(\sin\theta) \quad (17)$$

where  $A_i$  are the coefficients of the Legendre polynomials. To create the different artificial wind possibilities, the wind is then reconstructed as

$$U_{\text{rand}}^j(\theta) = \sum_{i=0}^{99} S_i^j A_i P_i(\sin\theta) \quad (18)$$

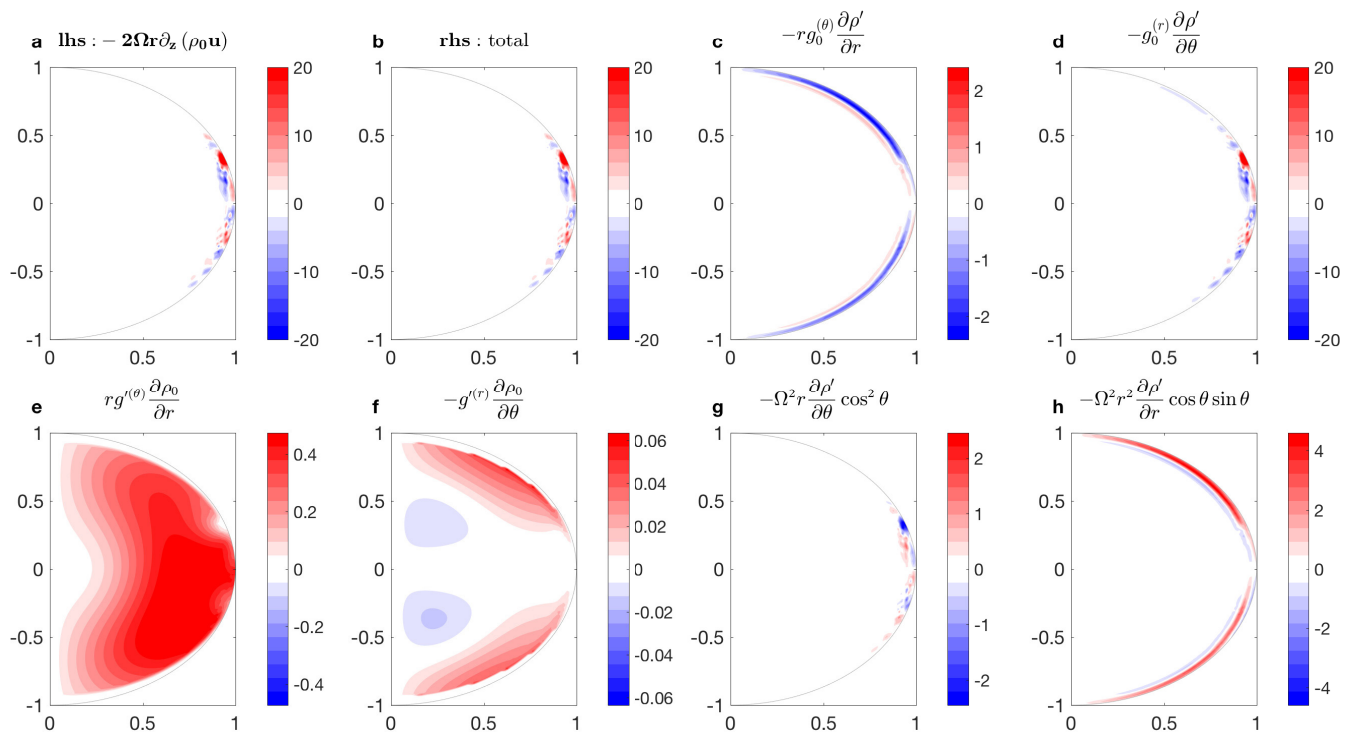
where  $S_i^j$  are a 100 plus or minus signs randomly chosen for each realization  $j$  of the wind. The resulting artificial cloud-level wind retains the basic characteristics (width and strength) of the observed zonal jets, but their latitudinal locations are now very different. To statistically examine our ability to reach a solution that gives a good match between the model-calculated gravity harmonics and those measured, we generated 1,000 artificial cloud-level wind profiles. A few examples of such randomly generated winds are shown in Extended Data Fig. 3. We note that while the wind profiles are very different from one another, the main characteristics of the observed winds are retained. Extended Data Fig. 4 shows the resulting  $J_3$ ,  $J_5$ ,  $J_7$  and  $J_9$  for these flow profiles (blue dots), optimized in the same way that the cloud-level wind solutions are. The results indicate that the gravity harmonics calculated using the specific cloud-level wind profile (black points with their uncertainty ellipse), give results closer to the measurements (red points with their uncertainty ellipse) than 99% of the random profiles, indicating the robustness of this result. We note the tendency of the optimized solutions to be in the quarter of the phase space where the measurements are (Extended Data Fig. 4), particularly for the case of  $J_5$  and  $J_7$ , because for these harmonics the absolute value of the measurement is largest and the relative measurement error is smallest (see Table 1), so their weight in the cost function is the largest. Taking the same random set of meridional profiles and calculating their gravity harmonics for a fixed vertical profile (without the optimization process) gives solutions spread equally over all quarters of the parameter space (Extended Data Fig. 5). This illustrates that the tendency of the simple exponential decay solution to have the correct sign and magnitude (Fig. 2) is also very likely not by chance. As an additional test we calculate the solution taking the Jupiter observed cloud-level meridional profile, but extended into the interior radially instead of along the direction of the spin axis. In this case even the sign of the gravity harmonics differs from the measurements. **Non-uniqueness of the gravity inversion.** It is important to note that the gravity inversion problem is non-unique, and as demonstrated in Figs. 3 and 4, different profiles can give similar gravity signatures. In addition, the cases presented here do not match the measurements perfectly, and with more free parameters and/or other meridional profiles<sup>45</sup> one could achieve better matches to the measurements. However, since the problem is non-unique, achieving a perfect match is not necessarily meaningful. Thus, the rationale of this study is to show that a minimal set of assumptions about the vertical and meridional structure gives by itself a very good, statistically significant, match to the measurements, indicating the structure and extent of the flow. Regardless of the exact vertical profile (which can depend on the parameterization and the non-uniqueness) the gravity measurements robustly reveal that the east–west jet streams on Jupiter are very deep, reaching several

thousands of kilometres beneath the cloud-level (several tens of kilobars of pressure), and advect a large mass that is on the order of one per cent of the mass of the planet.

**Code availability.** The code for inversion of the gravity measurements is available at [http://www.weizmann.ac.il/eserpages/kaspi/juno\\_code/](http://www.weizmann.ac.il/eserpages/kaspi/juno_code/).

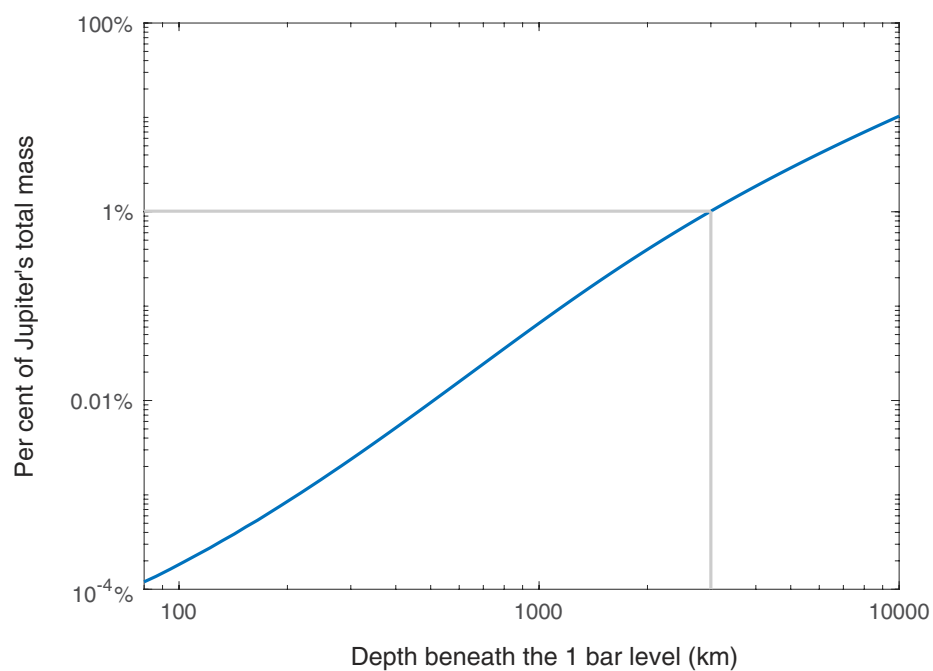
**Data availability.** Figure data are available upon request.

31. Hubbard, W. B. *Planetary Interiors* (Van Nostrand Reinhold, 1984).
32. Guillot, T. & Morel, P. CEPAM: a code for modeling the interiors of giant planets. *Astron. Astrophys. Suppl. Ser.* **109**, 109–123 (1995).
33. Militzer, B., Soubiran, F., Wahl, S. M. & Hubbard, W. B. Understanding Jupiter's interior. *J. Geophys. Res. Planets* **121**, 1552–1572 (2016).
34. Miguel, Y., Guillot, T. & Fayon, L. Jupiter internal structure: the effect of different equations of state. *Astron. Astrophys.* **596**, A114 (2016).
35. Helled, R. & Stevenson, D. J. The fuzziness of giant planets' cores. *Astrophys. J. Lett.* **840**, L4 (2017).
36. Wisdom, J. & Hubbard, W. B. Differential rotation in Jupiter: a comparison of methods. *Icarus* **267**, 315–322 (2016).
37. Pedlosky, J. *Geophysical Fluid Dynamics* (Springer, 1987).
38. Vallis, G. K. *Atmospheric and Oceanic Fluid Dynamics* (Cambridge Univ. Press, 2006).
39. Hubbard, W. B. Conventric Maclaurian spheroid models of rotating liquid planets. *Astrophys. J.* **768**, 43 (2013).
40. Kong, D., Zhang, K. & Schubert, G. Odd gravitational harmonics of Jupiter: effects of spherical versus nonspherical geometry and mathematical smoothing of the equatorially antisymmetric zonal winds across the equatorial plane. *Icarus* **277**, 416–423 (2016).
41. Folkner, W. M. *et al.* Jupiter gravity field from first two orbits by Juno. *Geophys. Res. Lett.* **44**, 4694–4700 (2017).
42. Porco, C. C. *et al.* Cassini imaging of Jupiter's atmosphere, satellites and rings. *Science* **299**, 1541–1547 (2003).
43. Kaspi, Y., Showman, A. P., Hubbard, W. B., Aharonson, O. & Helled, R. Atmospheric confinement of jet-streams on Uranus and Neptune. *Nature* **497**, 344–347 (2013).
44. Kong, D., Zhang, K. & Schubert, G. Wind-induced odd gravitational harmonics of Jupiter. *Mon. Not. R. Astron. Soc.* **450**, L11–L15 (2015).
45. Dowling, T. E. Estimate of Jupiter's deep zonal-wind profile from Shoemaker-Levy 9 data and Arnold's second stability criterion. *Icarus* **117**, 439–442 (1995).

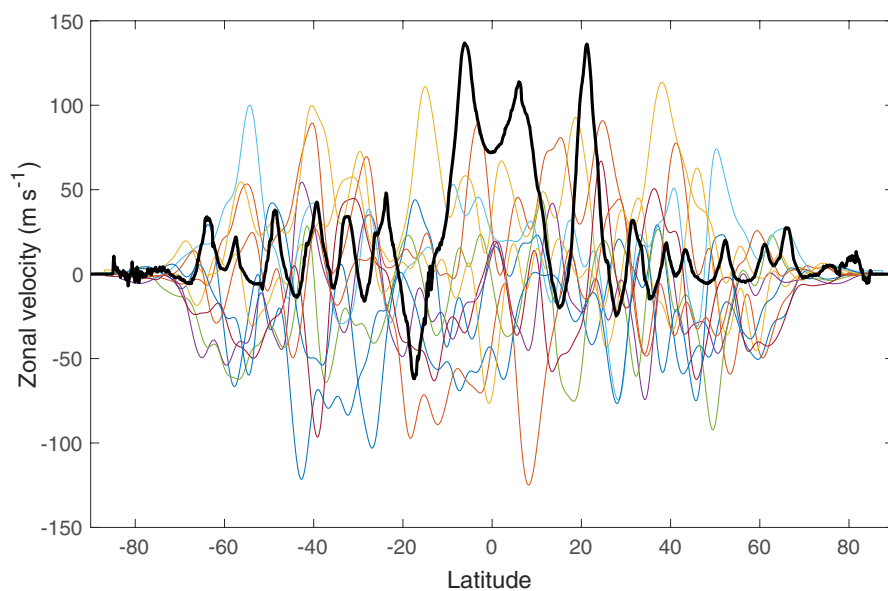


**Extended Data Figure 1 | The vorticity balance.** Solution to equation (6). **a**, Left-hand-side term with the wind profile from Fig. 3. **b**, Total of the right-hand side. **c–h**, The six terms on the right-hand side of equation (6), showing that the thermal wind balance (**a** and **d**) is the leading-order balance. Note that the different panels have different colour scales.

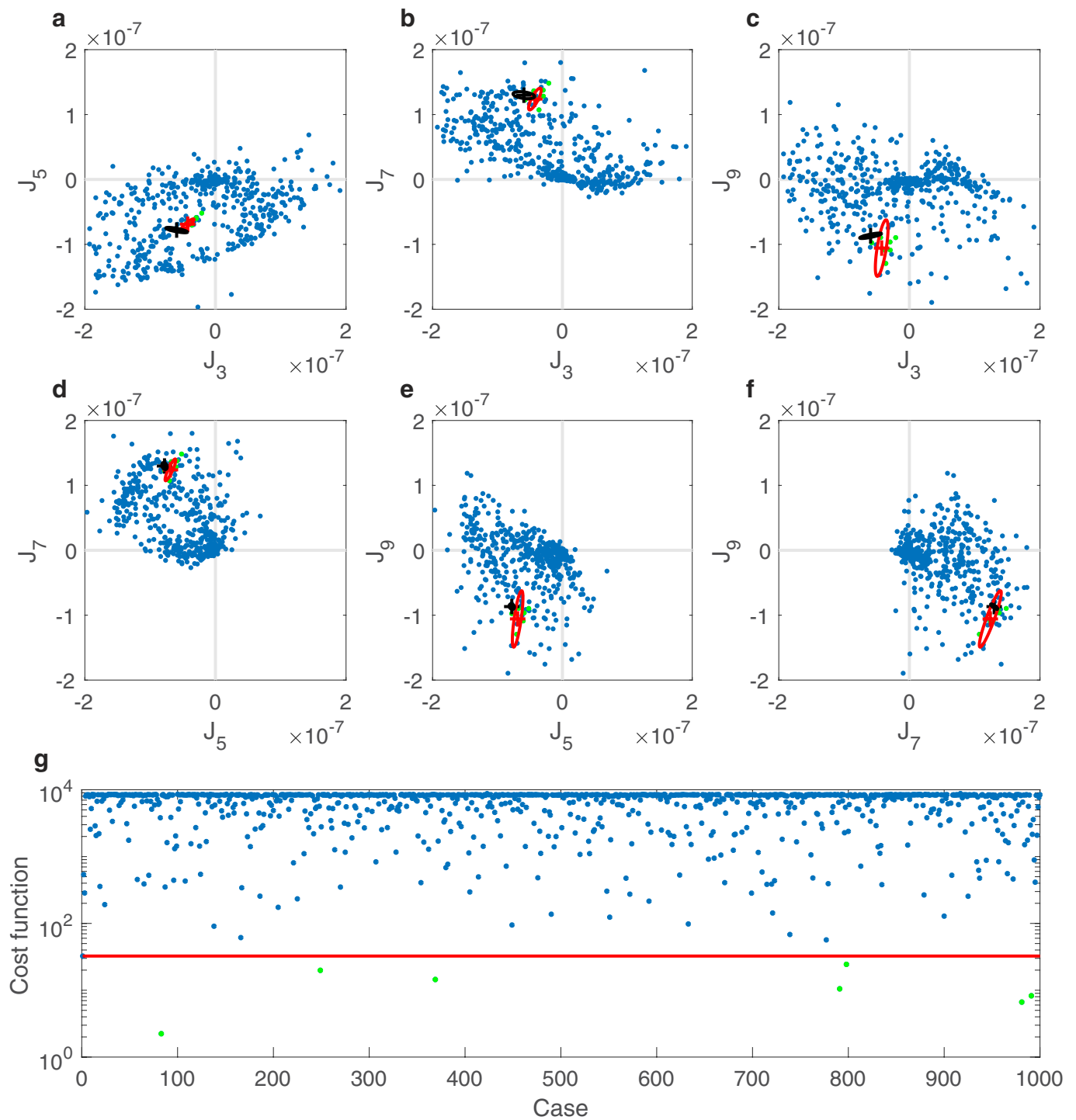




**Extended Data Figure 2 | Jupiter's mass distribution.** The percentage of Jupiter's mass as a function of depth beneath the 1-bar level. The grey line shows that roughly 1% of the mass is contained above a depth of 3,000 km.



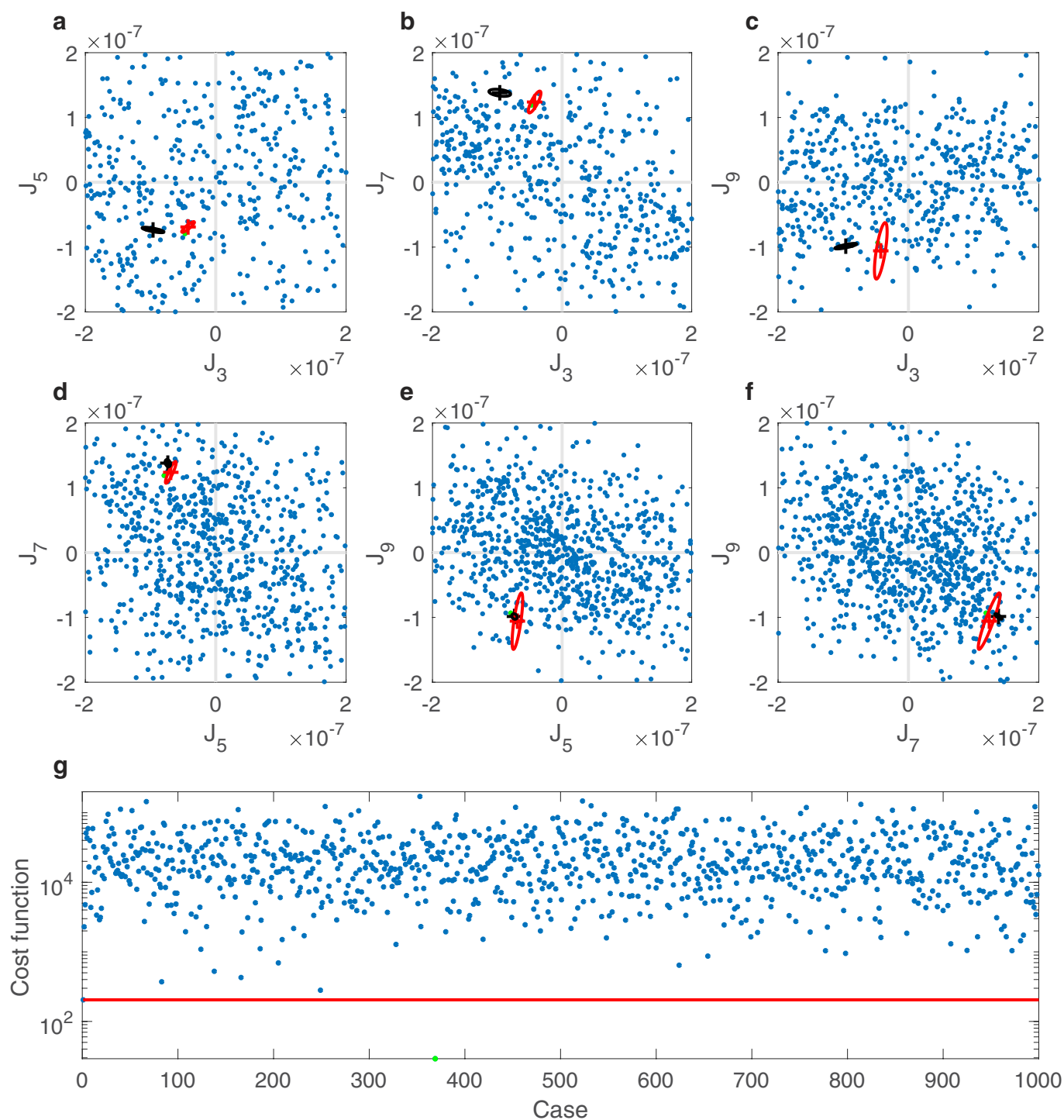
**Extended Data Figure 3 | Example of wind profiles used for the statistical significance test.** The observed cloud-level wind (black), together with a sample of ten randomly generated wind profiles.



**Extended Data Figure 4 | Optimized solutions for the odd harmonics using random zonal wind profiles.** a–f, Optimized solutions (blue) for  $J_3$ ,  $J_5$ ,  $J_7$  and  $J_9$  for flows with 1,000 different artificial meridional profiles of the zonal wind (as in Extended Data Fig. 3). The Juno measurements are shown in red with their corresponding uncertainty ellipse. The optimized solution corresponding to Jupiter's observed

cloud-level zonal wind profile (Fig. 3) is shown in black with the corresponding uncertainty ellipse. g, The cost function for all different meridional profiles explored, with the red line corresponding to the solution with the Jupiter zonal wind profile. Fewer than 1% of the solutions have lower cost functions (green).





**Extended Data Figure 5 | Solutions for the odd harmonics using random zonal wind profiles and a fixed vertical profile.** a–f, Solutions (blue) for  $J_3$ ,  $J_5$ ,  $J_7$  and  $J_9$  for flows with 1,000 different artificial meridional profiles of the zonal wind (as in Extended Data Fig. 3), and the vertical profile held fixed with  $H = 2,000$  km,  $\Delta H = 1,500$  km and  $\alpha = 1$ . The Juno measurements are shown in red with their corresponding uncertainty ellipse. The solution with these parameters and using Jupiter's observed cloud-level zonal wind profile is shown in black with the corresponding

uncertainty ellipse. g, The cost function for all different meridional profiles explored, with the red line corresponding to the solution with the Jupiter zonal wind profile. This shows that when no optimization is done (which takes into consideration the relative measurement error of the different harmonics), the solutions are spread equally over all four quadrants in these phase spaces (unlike in Extended Data Fig. 4). Only one solution has a lower cost function (green).

**Extended Data Table 1 | Flow-induced even gravity harmonics**

$\times 10^{-8}$	Model without latitudinal variation	Model with latitudinal variation
$\Delta J_2$	$54.62 \pm 5.21$	$-48.87 \pm 7.93$
$\Delta J_4$	$-5.18 \pm 0.74$	$-15.01 \pm 7.56$
$\Delta J_6$	$0.33 \pm 0.35$	$0.29 \pm 1.49$
$\Delta J_8$	$5.41 \pm 0.28$	$4.76 \pm 0.61$
$\Delta J_{10}$	$-5.35 \pm 0.25$	$-4.94 \pm 0.71$

The even gravity harmonics solutions for the optimization, with and without variation of flow depth with latitude, that correspond to the solutions presented in Figs 3 and 4 and Table 1. The uncertainties are the  $3\sigma$  uncertainty values.

**Extended Data Table 2 | The weights matrix  $W$  used in the cost function  $L$  of equation (16)**

	$J_3$	$J_5$	$J_7$	$J_9$
$J_3$	8.32	-11.05	1.45	-0. 41
$J_5$	-11.05	20.21	-12.26	3.35
$J_7$	1.45	-12.26	14.31	-7.63
$J_9$	-0. 41	3.35	-7.63	7.91

Shown are the weights associated with  $J_3$ ,  $J_5$ ,  $J_7$  and  $J_9$  (diagonal terms) and those associated with the correlation between the harmonics (off-diagonal terms). The values reflect the uncertainties in the measurements, calculated taking the inverse of the measurement error covariance matrix multiplied by 9 (to reflect  $3\sigma$  uncertainties). The larger the value, the larger the weight given to it when minimizing the cost function. Values shown are multiplied by  $10^{-16}$ .



# A suppression of differential rotation in Jupiter's deep interior

T. Guillot<sup>1</sup>, Y. Miguel<sup>1,2</sup>, B. Militzer<sup>3</sup>, W. B. Hubbard<sup>4</sup>, Y. Kaspi<sup>5</sup>, E. Galanti<sup>5</sup>, H. Cao<sup>6,7</sup>, R. Helled<sup>8</sup>, S. M. Wahl<sup>3</sup>, L. Iess<sup>9</sup>, W. M. Folkner<sup>10</sup>, D. J. Stevenson<sup>6</sup>, J. I. Lunine<sup>11</sup>, D. R. Reese<sup>12</sup>, A. Biekman<sup>1</sup>, M. Parisi<sup>10</sup>, D. Durante<sup>9</sup>, J. E. P. Connerney<sup>13</sup>, S. M. Levin<sup>10</sup> & S. J. Bolton<sup>14</sup>

**Jupiter's atmosphere is rotating differentially, with zones and belts rotating at speeds that differ by up to 100 metres per second. Whether this is also true of the gas giant's interior has been unknown<sup>1,2</sup>, limiting our ability to probe the structure and composition of the planet<sup>3,4</sup>. The discovery by the Juno spacecraft that Jupiter's gravity field is north–south asymmetric<sup>5</sup> and the determination of its non-zero odd gravitational harmonics  $J_3$ ,  $J_5$ ,  $J_7$  and  $J_9$  demonstrates that the observed zonal cloud flow must persist to a depth of about 3,000 kilometres from the cloud tops<sup>6</sup>. Here we report an analysis of Jupiter's even gravitational harmonics  $J_4$ ,  $J_6$ ,  $J_8$  and  $J_{10}$  as observed by Juno<sup>5</sup> and compared to the predictions of interior models. We find that the deep interior of the planet rotates nearly as a rigid body, with differential rotation decreasing by at least an order of magnitude compared to the atmosphere. Moreover, we find that the atmospheric zonal flow extends to more than 2,000 kilometres and to less than 3,500 kilometres, making it fully consistent with the constraints obtained independently from the odd gravitational harmonics. This depth corresponds to the point at which the electric conductivity becomes large and magnetic drag should suppress differential rotation<sup>7</sup>. Given that electric conductivity is dependent on planetary mass, we expect the outer, differentially rotating region to be at least three times deeper in Saturn and to be shallower in massive giant planets and brown dwarfs.**

Juno measurements of odd gravitational harmonics<sup>5</sup> constrain the maximum depth to which the observed atmospheric zonal flow persists<sup>6</sup>. These estimates, however, are based on the north–south asymmetries in the zonal flow, and cannot exclude the presence of a deeper north–south symmetric flow. Fortunately, further insights can be obtained by comparing the even gravitational harmonics obtained from interior models assuming rigid rotation with those expected for a differentially rotating planet. The harmonics from rigidly rotating interior models are highly correlated because they probe similar regions of the interior<sup>8</sup>. On the other hand, differential rotation similar to that observed in the cloud layer affects the different gravitational harmonics (moments) relatively evenly<sup>9,10</sup>.

We derive an ensemble of interior models with Jupiter's mass and equatorial radius using both the CEPAM code<sup>11</sup> and by perturbing density profiles obtained by the Concentric MacLaurin Spheroid (CMS) code<sup>12</sup>. Our range of  $J_2$  values is set by Juno's measurements and the maximum uncertainty due to the unknown interior differential rotation<sup>10</sup>. These models use different equations of state of hydrogen and helium<sup>13,14</sup>, including a possible jump of up to 500 K in temperature in the helium phase-separation region, and the possibility (or not) of a dilute core<sup>12</sup>. The calculation of the

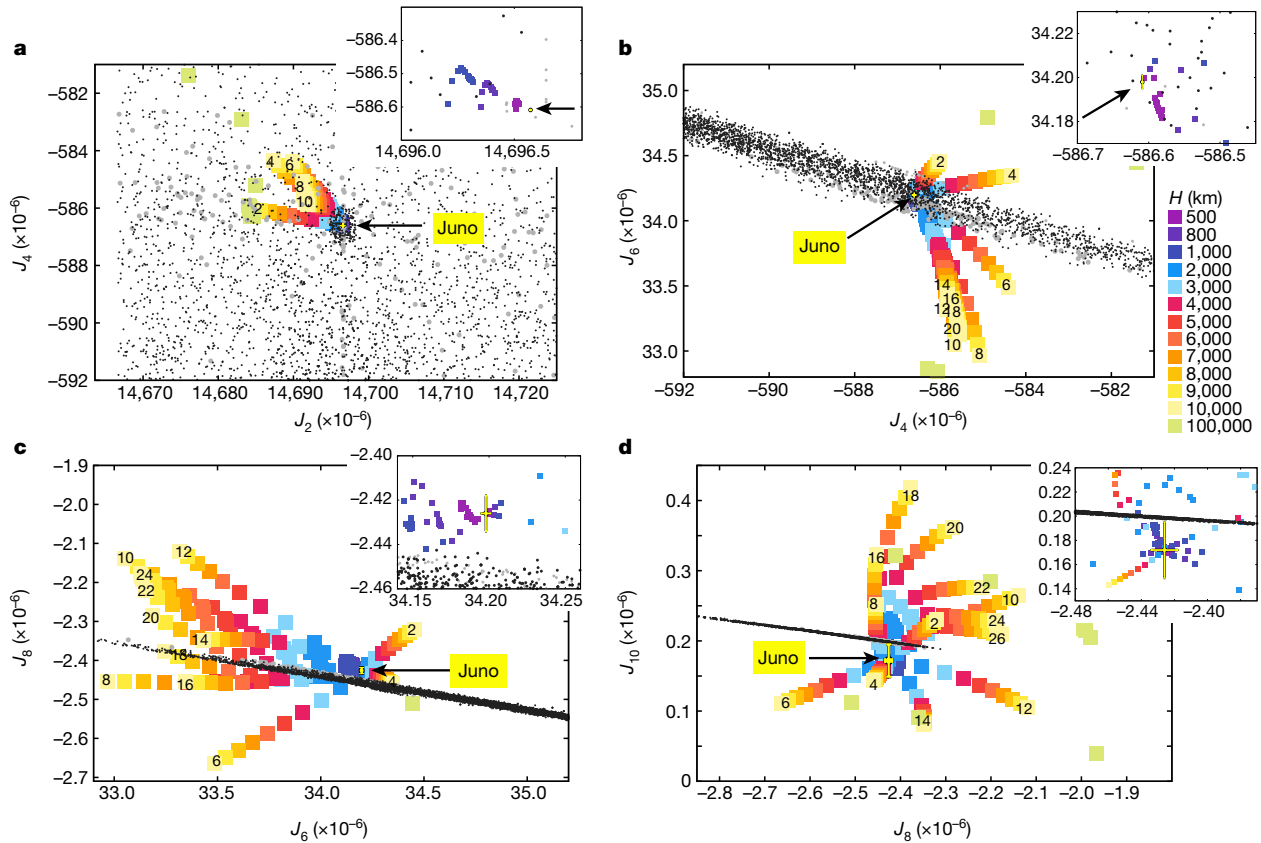
gravitational harmonics is performed in two ways, with the CMS theory<sup>15,16</sup> directly or with a fourth-order theory of figures<sup>17,18</sup> combined with a direct integration of the reconstructed two-dimensional density structure using a Gauss–Legendre quadrature. A calibration of the values obtained from the theory of figures to the CMS values ensures an accurate estimate of the high-order  $J$  values (see Methods).

The offset between differential and rigid rotation for each harmonic  $i$  (with  $2i = 2, 4, 6, 8, 10$ ),  $\Delta J_{2i} = J_{2i}^{\text{differential}} - J_{2i}^{\text{rigid}}$ , is calculated by assuming that the dynamical flows generate density perturbations that can be related through thermal wind balance<sup>10,19</sup>. We use a polynomial fit of degree  $m$  to the observed zonal winds<sup>20</sup> and an exponential decay in wind strength of  $e$ -folding depth  $H$ . We vary  $m$  between 2 and 30 and  $H$  between 0 km (rigid rotation) and 100,000 km (rotation on cylinders all the way to the centre of the planet), thus creating a wide range of possible interior flows. We use the Juno measurements<sup>5</sup> to calculate effective gravitational harmonics  $J_{2i}^{\text{eff}}(H, m) = J_{2i}^{\text{Juno}} - \Delta J_{2i}(H, m)$ . These are the values that must be matched by interior models assuming rigid rotation.

We compare the gravitational harmonics obtained from interior models to the effective gravitational harmonics in Fig. 1. Our interior models purposely cover a wide range of  $J_2$  values, compatible with the Juno measurements and variable interior differential rotation, varying from a solution representing a very shallow region with differential rotation at the surface to one representing a deep region extending to the planet's centre (Fig. 1a). We also allow for a wide range of meridional profiles ( $m$  values), to allow for the possibility that the internal flows have less latitudinal variation than the cloud-level wind profile. We see that the extent of interior model solutions is noticeably smaller in  $J_4$  versus  $J_6$  and becomes a well defined linear relation in  $J_6$  versus  $J_8$ , and  $J_8$  versus  $J_{10}$ . On the other hand, differential rotation affects the  $J_{2i}$  values more uniformly as a function of the parameters  $H$  and  $m$ . The solutions are obtained by matching rigidly rotating interior models (black and grey dots) to the effective gravitational harmonics (coloured squares).

In the  $J_2$  versus  $J_4$  plane, any value of the effective gravitational harmonics can be matched by small adjustments of the assumed interior composition: no constraint on interior differential rotation is possible. In the  $J_4$  versus  $J_6$ ,  $J_6$  versus  $J_8$ , and  $J_8$  versus  $J_{10}$  planes, the same interior models are incompatible with most values of the effective gravitational harmonics. The corresponding values of  $H$  and  $m$  are therefore excluded. In the  $J_4$  versus  $J_6$  plane, the interior models cross the Juno point, providing only an upper limit on  $H$ . However, in the  $J_6$  versus  $J_8$ , and  $J_8$  versus  $J_{10}$  planes, the slight offset between the Juno point and the interior model area implies that a lower limit on  $H$  may be derived.

<sup>1</sup>Université Côte d'Azur, OCA, Lagrange CNRS, 06304 Nice, France. <sup>2</sup>Leiden Observatory, University of Leiden, Niels Bohrweg 2, 2333CA Leiden, The Netherlands. <sup>3</sup>University of California, Berkeley, California 94720, USA. <sup>4</sup>Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA. <sup>5</sup>Weizmann Institute of Science, Rehovot 76100, Israel. <sup>6</sup>California Institute of Technology, Pasadena, California 91125, USA. <sup>7</sup>Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA. <sup>8</sup>University of Zurich, 8057 Zurich, Switzerland. <sup>9</sup>Sapienza Università di Roma, 00184 Rome, Italy. <sup>10</sup>Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. <sup>11</sup>Cornell University, Ithaca, New York 14853, USA. <sup>12</sup>LESIA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Universités, UPMC Université Paris 06, Université Paris Diderot, Sorbonne Paris Cité, 92195 Meudon, France. <sup>13</sup>NASA/GSFC, Greenbelt, Maryland, USA. <sup>14</sup>Southwest Research Institute, San Antonio, Texas, USA.

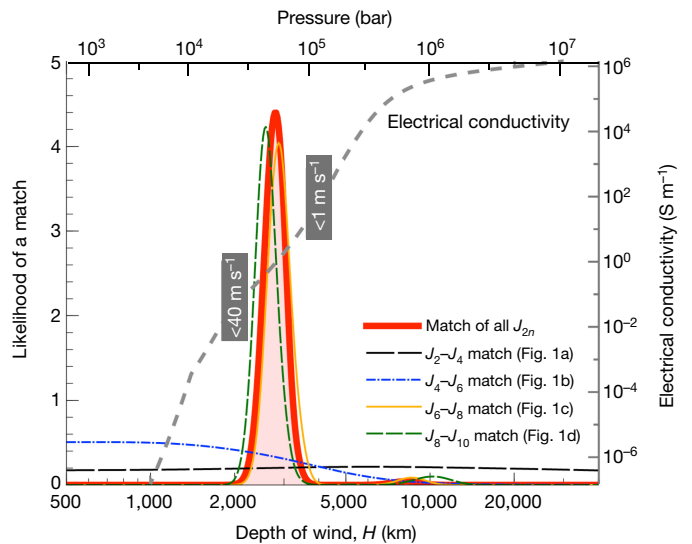


**Figure 1 | Jupiter's gravitational harmonics  $J_2$  to  $J_{10}$ .** **a**,  $J_2$  versus  $J_4$ . **b**,  $J_4$  versus  $J_6$ . **c**,  $J_6$  versus  $J_8$ . **d**,  $J_8$  versus  $J_{10}$ . The points correspond to interior models of Jupiter calculated assuming rigid rotation using CEPAM<sup>11</sup> (black points) and CMS<sup>12,15</sup> (grey points). The coloured squares correspond to the values that must be matched by interior models in order to be considered successful solutions for observed zonal flows extending

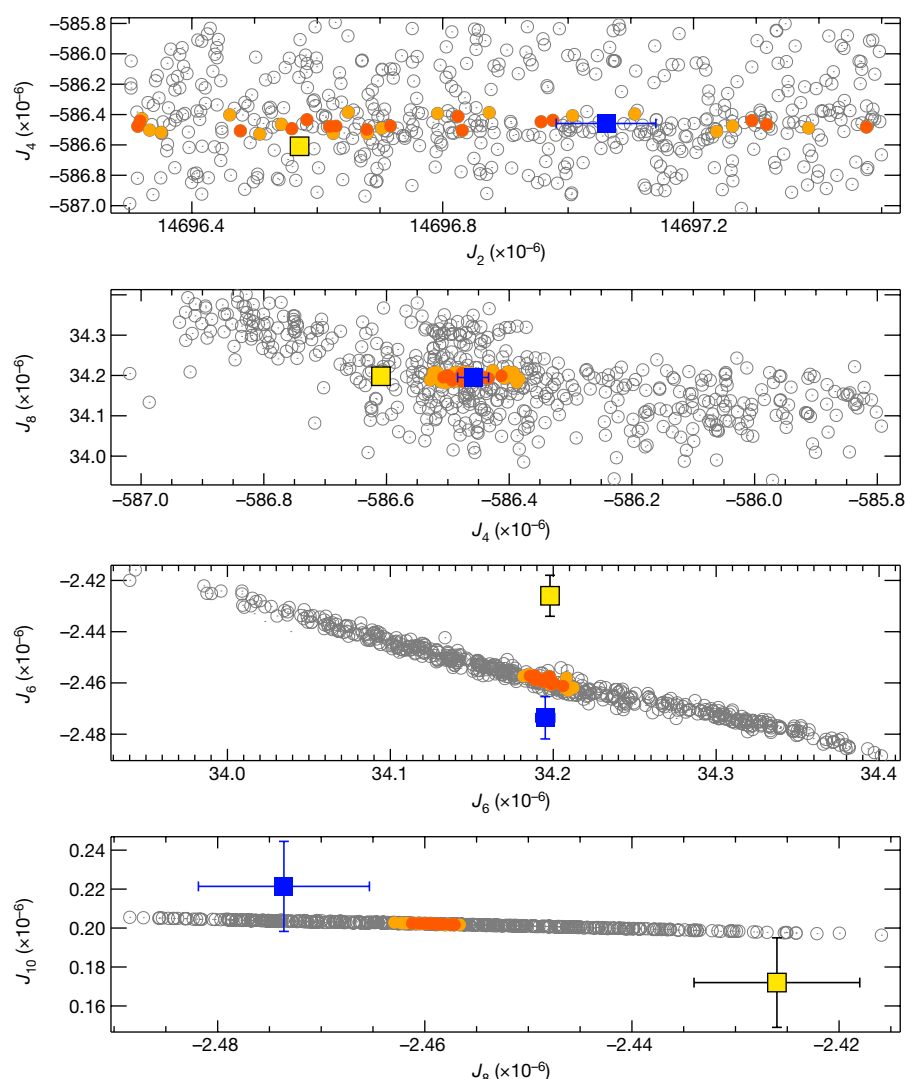
to various depths, from  $H = 500$  km to  $H = 100,000$  km, and by filtering the atmospheric flow ( $m$  from 2 to 30; see text)<sup>10</sup>. The numbers on the plots correspond to the value of  $m$  for  $H = 10,000$  km. The Juno measurements and their  $1\sigma$  error bars are shown in yellow. Because these are extremely small, arrows point to the corresponding points. Insets are close-ups around the Juno points for all four panels.

Mechanisms other than differential rotation cannot realistically explain that offset: in order to alter the relations between  $J_6$ ,  $J_8$  and  $J_{10}$ , they would need to strongly affect the interior density profile in the outer approximately 30% of the planet<sup>8</sup>. In this region, uncertainties in the H–He phase separation and related composition jumps are included in the interior model and constrained by the  $J_4$  versus  $J_6$  values. The other source of uncertainty is related to the condensation of water and silicates but is expected to affect  $J_4$  by only about  $10^{-7}$ ,  $J_6$  by  $10^{-8}$  and  $J_8$  by about  $10^{-9}$ , that is, more than one order of magnitude less than required (see Methods).

To estimate possible values of the wind depth  $H$  (measured from the 1-bar level, approximately the cloud tops), we calculate the likelihood that an atmospheric model (accounting for the effect of differential rotation) combined with an interior model (accounting for the effect of interior structure) matches the observed even gravity coefficients. For a given value of  $H$ , we integrate the function  $\exp[-(J_{2i}^{\text{eff}}(H, m) - J_{2i}^{\text{model}})^2 / (2\sigma_{2i}^2)] / [(2\pi)^{1/2} \sigma_{2i}]$  over all models in our ensemble and all values of  $m$ .  $\sigma_{2i}$  encompasses the  $1\sigma$  uncertainty of the Juno measurements as well as the variance in our ensemble of models. Figure 2 confirms the analysis of Fig. 1 that  $J_2$  versus  $J_4$  or  $J_4$  versus  $J_6$  alone cannot be used to constrain the wind depth  $H$ . The strongest constraints on  $H$  come from the  $J_6$  versus  $J_8$  and  $J_8$  versus  $J_{10}$  planes because the weights of atmospheric contributions become large relative to those for the lower harmonics. When constraints from  $J_2$  to  $J_{10}$  are combined, a strong peak emerges in the likelihood function in Fig. 2. Only values of  $H$  between 2,000 km and 3,500 km are compatible with the available data. This depth corresponds to the one at which the electrical conductivity<sup>21</sup> increases to a modest value ( $0.01$ – $1$  S m<sup>-1</sup>) and



**Figure 2 | Constraint on the depth  $H$  of Jupiter's zonal flow obtained from interior models and Juno's even gravitational harmonics.** The lines correspond to Fig. 1a–d:  $J_2$  versus  $J_4$ ,  $J_4$  versus  $J_6$ ,  $J_6$  versus  $J_8$ , and  $J_8$  versus  $J_{10}$ . The profile of electrical conductivity in Jupiter's interior<sup>21</sup> is shown for comparison. Ohmic dissipation is expected to limit zonal flows<sup>7</sup> to less than  $40$  m s<sup>-1</sup> at a depth of 2,000 km and to  $1$  m s<sup>-1</sup> at 4,000 km. Only interior models with  $-586.8 < J_4 \times 10^6 < -584.5$  (corresponding to the maximum range of  $J_4^{\text{eff}}$  values allowed by differential rotation) were included in the calculation.



**Figure 3 | Ensemble of interior models of Jupiter fitting the even gravitational harmonics  $J_2$  to  $J_{10}$ .** The Juno values are shown as yellow squares with  $1\sigma$  error bars. The blue squares with  $1\sigma$  error bars correspond to the effective gravitational harmonics obtained when accounting for the

differential rotation derived from Jupiter's odd gravitational harmonics<sup>6</sup>. Interior models fitting all effective gravitational harmonics  $J_4$  to  $J_{10}$  (blue squares) are highlighted in colour depending on whether they fit within  $2\sigma$  (dark orange) or  $3\sigma$  (light orange).

the Lorentz force associated with the zonal flow (magnetic drag) becomes comparable to the observed divergence of the Reynolds stress in the cloud layers<sup>7,22,23</sup>. Indeed, energy budget considerations of the ohmic dissipation being smaller than the observed luminosity predict a penetration depth between about 2,000 km and 2,800 km below the cloud tops of Jupiter<sup>7,24</sup>.

The results obtained in Figs 1 and 2 are based on a simple law (an exponential decay of the atmospheric zonal flow) that was obtained independently of Juno's measurements<sup>10</sup>. In Fig. 3 we show that the more elaborate differential-rotation law that is fitted to Jupiter's odd gravitational harmonics<sup>6</sup> is consistent with the interior models, confirming that the symmetric and asymmetric parts of the observed zonal flow extend to a similar depth. The solutions matching the observations generally cover an extensive parameter space (see Extended Data Table 1). One salient feature is that these solutions are characterized by an increase of the heavy-element abundance in the deeper interior, either where hydrogen becomes metallic or deeper in a dilute core, confirming the results obtained after Juno's first two orbits<sup>12</sup>.

Furthermore, by adopting the differential rotation law for the upper 3,000 km of Jupiter's atmosphere, we can provide approximate constraints on the rotation of the deeper parts of the planet. To do so, we assume that the deeper interior rotates on cylinders all the way to the centre and adopt a scaled version of the  $\Delta J_{2i}$  relations from Fig. 1. We

calculate the likelihood of such a model with unknown deep differential rotation  $\nu$  between zero and the observed atmospheric rotation of about  $100 \text{ m s}^{-1}$ , using the same approach as for Fig. 2 (see Methods). The results are shown in Extended Data Fig. 2. Only an upper limit may be derived on  $\nu$ : beneath the first 3,000-km-deep layer, deep differential rotation must be limited to amplitudes at least an order of magnitude smaller than the observed atmospheric ones.

The observed winds thus penetrate deep in the atmosphere all the way to the levels at which the conductivity and the resulting magnetic drag become large enough to force fluid motions into rigid-body rotation<sup>23,24</sup>. In gaseous planets, electrical conductivity strongly increases with pressure, which is itself a strong function of the planetary mass. In Saturn, one must go three times deeper than in Jupiter to reach the same conductivity<sup>7,21</sup>. Saturn has a similar intrinsic luminosity but a magnetic field that is an order of magnitude smaller than Jupiter's<sup>25</sup>. We hence expect Saturn's outer, differentially rotating region to extend to at least 9,000 km, which should leave a strong imprint on its gravity field. Conversely, massive giant exoplanets and brown dwarfs should have shallower differentially rotating, outer envelopes<sup>26</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.



Received 19 September 2017; accepted 17 January 2018.

- Busse, H. F. A simple model of convection in the Jovian atmosphere. *Icarus* **29**, 255–260 (1976).
- Vasavada, A. R. & Showman, A. P. Jovian atmospheric dynamics: an update after Galileo and Cassini. *Rep. Prog. Phys.* **68**, 1935–1996 (2005).
- Hubbard, W. B. Effects of differential rotation on the gravitational figures of Jupiter and Saturn. *Icarus* **52**, 509–515 (1982).
- Guillot, T., Gautier, D. & Hubbard, W. B. New constraints on the composition of Jupiter from Galileo measurements and interior models. *Icarus* **130**, 534–539 (1997).
- Iess, L. *et al.* Measurement of Jupiter's asymmetric gravity field. *Nature* **555**, <https://doi.org/10.1038/nature25776> (2018).
- Kaspi, Y. *et al.* Jupiter's atmospheric jet streams extend thousands of kilometres deep. *Nature* **555**, <https://doi.org/10.1038/nature25793> (2018).
- Cao, H. & Stevenson, D. J. Zonal flow magnetic field interaction in the semi-conducting region of giant planets. *Icarus* **296**, 59–72 (2017).
- Guillot, T. The interiors of giant planets: models and outstanding questions. *Annu. Rev. Earth Planet. Sci.* **33**, 493–530 (2005).
- Hubbard, W. B. Gravitational signature of Jupiter's deep zonal flows. *Icarus* **137**, 357–359 (1999).
- Kaspi, Y. *et al.* The effect of differential rotation on Jupiter's low-degree even gravity moments. *Geophys. Res. Lett.* **44**, 5960–5968 (2017).
- Miguel, Y., Guillot, T. & Fayon, L. Jupiter internal structure: the effect of different equations of state. *Astron. Astrophys.* **596**, A114 (2016).
- Wahl, S. M. *et al.* Comparing Jupiter interior structure models to Juno gravity measurements and the role of an expanded core. *Geophys. Res. Lett.* **44**, 4649–4659 (2017).
- Militzer, B. & Hubbard, W. B. Ab initio equation of state for hydrogen-helium mixtures with recalibration of the giant-planet mass-radius relation. *Astrophys. J.* **774**, 148 (2013).
- Becker, A. *et al.* Ab initio equations of state for hydrogen (H-REOS.3) and helium (He-REOS.3) and their implications for the interior of brown dwarfs. *Astrophys. J. Suppl. Ser.* **215**, 21 (2014).
- Hubbard, W. B. Concentric Maclaurin spheroid models of rotating liquid planets. *Astrophys. J.* **768**, 43 (2013).
- Wisdom, J. & Hubbard, W. B. Differential rotation in Jupiter: a comparison of methods. *Icarus* **267**, 315–322 (2016).
- Zharkov, V. N. & Trubitsyn, V. P. *Physics Of Planetary Interiors* (Astronomy and Astrophysics Series, Pachart, 1978).
- Nettelmann, N. Low- and high-order gravitational harmonics of rigidly rotating Jupiter. *Astron. Astrophys.* **606**, A139 (2017).
- Kaspi, Y., Showman, A. P., Hubbard, W. B., Aharonson, O. & Helled, R. Atmospheric confinement of jet streams on Uranus and Neptune. *Nature* **497**, 344–347 (2013).
- Ingersoll, A. P. *et al.* in *Jupiter. The Planet, Satellites And Magnetosphere* (eds Bagenal, F., Dowling, T. E. & McKinnon, W. B.) *Cambridge Planetary Science Vol. 1*, 105–128 (Cambridge Univ. Press, 2004).
- French, M. *et al.* Ab initio simulations for material properties along the Jupiter adiabat. *Astrophys. J. Suppl. Ser.* **202**, 5 (2012).
- Salyk, C., Ingersoll, A. P., Lorre, J., Vasavada, A. & Del Genio, A. D. Interaction between eddies and mean flow in Jupiter's atmosphere: analysis of Cassini imaging data. *Icarus* **185**, 430–442 (2006).
- Schneider, T. & Liu, J. Formation of jets and equatorial superrotation on Jupiter. *J. Atmos. Sci.* **66**, 579–601 (2009).
- Liu, J., Goldreich, P. M. & Stevenson, D. J. Constraints on deep-seated zonal winds inside Jupiter and Saturn. *Icarus* **196**, 653–664 (2008).
- Connerney, J. E. P. in *Planets and Satellites* (eds Schubert, G. & Spohn, T.) *Treatise in Geophysics Vol. 10.06*, 195–237 (Elsevier, 2015).
- Showman, A. P. & Guillot, T. Atmospheric circulation and tides of “51 Pegasis b-like” planets. *Astron. Astrophys.* **385**, 166–180 (2002).

**Acknowledgements** This research was carried out at the Observatoire de la Côte d'Azur under the sponsorship of the Centre National d'Etudes Spatiales; at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA; by the Southwest Research Institute under contract with NASA; and at the Weizmann Institute of Science under contract with the Israeli Space Agency. Computations were performed on the ‘Mesocentre SIGAMM’ machine, hosted by the Observatoire de la Côte d'Azur.

**Author Contributions** T.G., Y.M. and B.M. ran interior models of Jupiter and carried out the analysis. W.B.H. and A.B. compared gravitational harmonics obtained by different methods. E.G. and Y.K. calculated the offset introduced by differential rotation. H.C., R.H., D.J.S. and J.I.L. provided theoretical support. S.M.W. provided additional interior models of Jupiter. D.R.R. provided a routine to calculate high-order gravitational harmonics efficiently. W.M.F., M.P. and D.D. carried out the analysis of the Juno gravity data. J.E.P.C., S.M.L. and S.J.B. supervised the planning, execution and definition of the Juno gravity experiment.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to T.G. ([tristan.guillot@oca.eu](mailto:tristan.guillot@oca.eu)).

**Reviewer Information** *Nature* thanks J. Fortney and N. Nettelmann for their contribution to the peer review of this work.

## METHODS

**Calculation of interior models.** The internal structure of Jupiter is calculated using the equations of hydrostatic equilibrium, energy transport, energy and mass conservation, which are solved with the interior structure code CEPAM<sup>27</sup>. These models are constructed to fit observational constraints such as Jupiter's radius and gravitational harmonics.

We adopt a four-layer structure for the interior models: (1) a helium-poor upper envelope in which hydrogen is in molecular form, (2) a helium-rich, metallic-hydrogen lower envelope, (3) a dilute core which consists of helium-rich metallic hydrogen with an increase of the heavy-element content and (4) a central dense core of ices and rocks. Because convection tends to homogenize large fractions of the envelope<sup>28</sup>, we expect that regions (1) and (2) should be largely convective and homogeneous. However, the presence of a phase separation<sup>29</sup> of helium in metallic hydrogen at about 1 Mbar may create a barrier to convection<sup>30–32</sup> and thus yield an increase in both helium and heavy-element abundances. The dilute core region may be inhomogeneous and an extension of the core itself<sup>33,34</sup>.

The determination of Jupiter's internal structure still rests on the accuracy of the equations of state<sup>11,35,36</sup>. For H and He we use two of the most recently published equations of state calculated from *ab initio* simulations: MH13<sup>13</sup> and REOS3<sup>14</sup>. For REOS3-H and REOS3-He, the pure hydrogen and pure helium equation-of-state tables, respectively, we calculate the entropy with a dedicated procedure<sup>11</sup>. MH13 was produced for a fixed mixture of H and He. To allow different concentrations of H and He in the different layers we extract from MH13 the table for H and since MH13 does not cover the entire pressure range in Jupiter's interior we merge the extracted table with the Saumon–Chabrier–van Horn equation of state<sup>11,36</sup>. The heavy elements are assumed to be composed of rocks and ices<sup>37</sup>.

Since we attempt to calculate the largest possible ensemble of realistic interior models we allow for the possibility of either efficient convection or double-diffusive convection in the helium phase-separation region by including a possible jump in temperature in that region<sup>30–32,35,38</sup>. Uncertainties in the location and characteristics of the helium phase-separation zone are considered by varying the limit<sup>29</sup> between region (1) and region (2) between 0.8 Mbar and 3 Mbar. Uncertainty about the presence of the dilute-core region (3) is included by performing some of the calculations either without this region (three-layer models) or with region (3) and including three variable parameters: the location of the transition, its smoothness and the heavy-element fraction in the transition region.

To obtain this large ensemble of possible interior models, for each set of imposed parameters, we obtain the mass fraction of ices in region (1) and the core mass that best fits the observed equatorial radius of the Jupiter<sup>39</sup>,  $71,492 \pm 4$  km and the gravitational harmonic  $J_2$  following an optimization procedure<sup>40</sup>. We do not attempt to fit the other gravitational harmonics and we allow for a large range of values for  $J_2$  between 0.014665 and 0.014725 in order to probe the ensemble of possible solutions, from rigidly rotating solutions to differential rotation extending all the way to the planetary centre.

Extended Data Table 1 summarizes the parameters used in the models. Their values are drawn either from a Gaussian distribution when they are constrained observationally or from a uniform distribution when we do not have sufficient a priori knowledge of their value. More than 200,000 interior models were calculated.

We calculate models in which the amount of water and rocks is suppressed at temperatures below 200 K and 3,000 K, respectively, in order to mimic the condensation of these species. The changes to  $J_4$  (about  $10^{-7}$ ), to  $J_6$  (about  $10^{-8}$ ) and to  $J_8$  (about  $10^{-9}$ ) are found to be too small to affect the results.

We also use an alternative method in which we perturb the density profiles for Jupiter<sup>9</sup> and calculate their gravitational harmonics using CMS. We introduce between 1 and 4 density jumps at random pressures. The magnitudes of the density changes are also chosen randomly between  $-5\%$  and  $+5\%$  to represent possible compositional deviations or equation-of-state deviations that are not yet understood. These thus represent a wide ensemble of models—some of them unphysical (for example, because of a decrease in density with increasing pressure). Nevertheless, the inferred ensemble of gravitational harmonics (grey points in Fig. 1) overlaps very closely with that obtained using full interior structure models (black points), suggesting that the results, in terms of the gravitational moments of a rigidly rotating Jupiter, are robust.

**Calculation of gravitational harmonics.** The calculation of the gravitational harmonics is performed as follows: for the CMS model and their perturbations we use the CMS approach<sup>15,16</sup>. For the CEPAM models, we use the faster theory of figures to fourth order<sup>17,18</sup> to obtain a bi-dimensional interior density profile  $\rho(\zeta, \theta)$  where  $\zeta$  is the (dimensionless) mean radius and  $\theta$  the colatitude. We then calculate the gravitational harmonics  $J_l$  as:

$$J_l = -\frac{1}{MR^l} \int_0^1 \int_0^{2\pi} \int_0^\pi r^l \rho(\zeta, \theta) P_l(\cos\theta) r^2 |r_\zeta| \sin\theta d\theta d\phi d\zeta$$

where  $M$  and  $R$  are the planetary mass and equatorial radius, respectively,  $r_\zeta$  is the partial derivative of  $r$  with respect to  $\zeta$ , and  $P_l(\cos\theta)$  is the Legendre polynomial of degree  $l$ . We use a Gauss–Legendre quadrature in the horizontal direction  $\theta$  and finite differences in the radial direction  $\zeta$ .

Extended Data Table 2 shows a comparison of solutions obtained from this method and from two other approaches. First, we use CEPAM on an  $n = 1$  polytropic equation of state and compare the solution to that calculated using an extremely accurate method<sup>16</sup>. The results are in good agreement, with offsets being at most  $1.5 \times 10^{-7}$ . These offsets are a natural consequence of the theory of figures expansion<sup>17,18</sup>. We then compare more realistic Jupiter models calculated with CEPAM and with the CMS method. The offsets for high-order harmonics are remarkably similar to the ones obtained for the polytropic model. The offsets for  $J_2$  are comparatively more important and are believed to be due to discretization errors<sup>16</sup>. These imply a small error on the core mass and the mass of heavy elements in the planet by an amount that is negligible in regard to the other uncertainties<sup>18</sup>. By comparing the solutions obtained with two slightly different models having the same  $J_2$  value with CEPAM and CMS, respectively (line REOS1a–1b in Extended Data Table 2), we can see that the offset in  $J_2$  has a small effect on  $J_4$  and an even smaller one on higher-order harmonics.

Using these results, we adopt the following offsets  $\delta J_4 = 0.11 \times 10^{-6}$ ,  $\delta J_6 = -0.057 \times 10^{-6}$ ,  $\delta J_8 = 0.166 \times 10^{-6}$ ,  $\delta J_{10} = -0.029 \times 10^{-6}$ . Although we expect this offset to change slightly as a function of the parameters used, the level of precision obtained is sufficient to derive constraints on the internal differential rotation. This is shown in Extended Data Fig. 1, which compares calculations performed with the different approaches.

**Constraints on deep differential rotation.** To derive constraints on the amount of differential rotation underneath the 'atmospheric' layer, we proceed as follows: First we imagine that we can divide the interior into a differentially rotating outer shell tied to the atmospheric zonal wind and a deeper layer with a smaller amount of differential rotation (with characteristic zonal velocity  $v$ ) all the way to the centre of Jupiter. Given that the rotation of the outer shell is constrained by the odd harmonics, we wish to find the possible values of  $v$ . We therefore need to associate effective gravitational harmonics  $J_{2l}^{\text{obs}}$  with each value of  $v$ .

We do so by adding Juno's value, the offset derived from the latitude-dependent flow profile that best fits Juno's odd harmonics, and a deeper component that is obtained from the purely cylindrical component for  $H = 100,000$  km (see Fig. 1)<sup>10</sup>:

$$J_{2l}^{\text{obs}} = J_{2l}^{\text{Juno}} - \delta J_{2l}^{\text{offset}} - \frac{v}{100 \text{ m s}^{-1}} \delta J_{2l}^{H=100,000 \text{ km}}(m)$$

where we assume that the value of  $\delta J_{2l}$  obtained for the atmospheric zonal flows ( $v \approx 100 \text{ m s}^{-1}$ ) may be scaled linearly for any characteristic velocity  $v$ .

We then calculate the likelihood of these models as a function of  $v$  with the same approach as for Fig. 2, including all gravitational harmonics  $J_4$  to  $J_{10}$ . The results are plotted in Extended Data Fig. 2. For our preferred model, we obtain a strong upper limit at  $10 \text{ m s}^{-1}$  with a preference for smaller values of  $v$ . For  $v < 6 \text{ m s}^{-1}$  the best interior models are found to lie within two standard deviations of all effective gravitational harmonics. For comparison, a model with a thin weather layer ( $H = 0$ ) and differential rotation on cylinders to the centre with velocity  $v$  is also found to favour small values of  $v < 10 \text{ m s}^{-1}$  but is incompatible with Juno's gravitational harmonics.

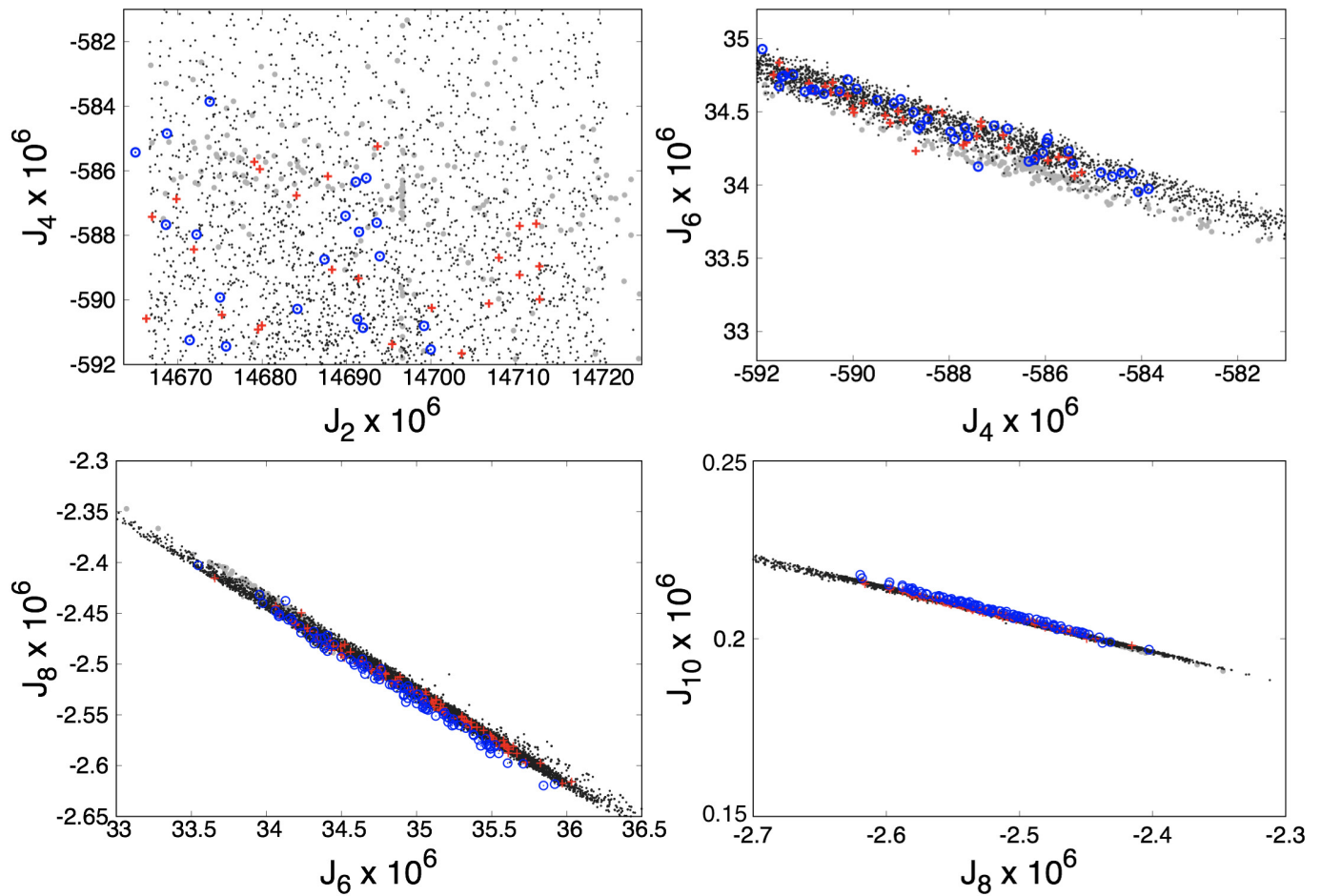
**Code availability.** The CEPAM code is available for download at <https://svn.oca.eu/codes/CEPAM/trunk>.

**Data availability.** Data sharing is not applicable to this article as no datasets were generated during the current study.

27. Guillot, T. & Morel, P. CEPAM: a code for modeling the interiors of giant planets. *Astron. Astrophys.* **109**, 109–123 (1995).
28. Vazan, A., Helled, R., Podolak, M. & Kovetz, A. The evolution and internal structure of Jupiter and Saturn with compositional gradients. *Astrophys. J.* **829**, 118 (2016).
29. Morales, M. A., Hamel, S., Caspersen, K. & Schwegler, E. Hydrogen-helium demixing from first principles: from diamond anvil cells to planetary interiors. *Phys. Rev. B* **87**, 174105 (2013).
30. Stevenson, D. J. & Salpeter, E. E. The dynamics and helium distribution in hydrogen-helium fluid planets. *Astrophys. J. Suppl. Ser.* **35**, 239–261 (1977).
31. Nettelmann, N., Fortney, J. J., Moore, K. & Mankovich, C. An exploration of double diffusive convection in Jupiter as a result of hydrogen-helium phase separation. *Mon. Not. R. Astron. Soc.* **447**, 3422–3441 (2015).
32. Mankovich, C., Fortney, J. J. & Moore, K. L. Bayesian evolution models for Jupiter with helium rain and double-diffusive convection. *Astrophys. J.* **832**, 113 (2016).
33. Stevenson, D. J. Cosmochemistry and structure of the giant planets and their satellites. *Icarus* **62**, 4–15 (1985).
34. Helled, R. & Stevenson, D. The fuzziness of giant planets' cores. *Astrophys. J.* **840**, L4 (2017).

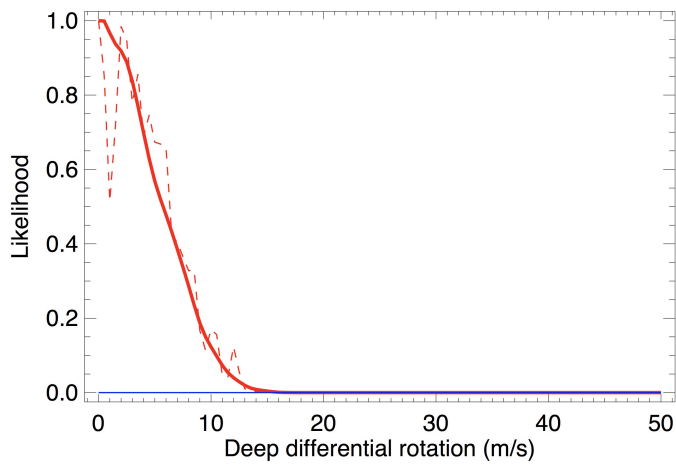
35. Hubbard, W. B. & Militzer, B. A preliminary Jupiter model. *Astrophys. J.* **820**, 80 (2016).
36. Saumon, D., Chabrier, G. & van Horn, H. M. An equation of state for low-mass stars and giant planets. *Astrophys. J. Suppl. Ser.* **99**, 713–741 (1995).
37. Saumon, D. & Guillot, T. Shock compression of deuterium and the interiors of Jupiter and Saturn. *Astrophys. J.* **609**, 1170–1180 (2004).
38. Leconte, J. & Chabrier, G. A new vision of giant planet interiors: impact of double diffusive convection. *Astron. Astrophys.* **540**, A20 (2012).
39. Lindal, G. F. The atmosphere of Neptune—an analysis of radio occultation data acquired with Voyager 2. *Astron. J.* **103**, 967–982 (1992).
40. Guillot, T. A comparison of the interiors of Jupiter and Saturn. *Planet. Space Sci.* **47**, 1183–1200 (1999).
41. Seiff, A. *et al.* Thermal structure of Jupiter's atmosphere near the edge of a 5-micron hot spot in the north equatorial belt. *J. Geophys. Res.* **103**, 22857–22889 (1998).
42. Serenelli, A. M. & Basu, S. Determining the initial helium abundance of the Sun. *Astrophys. J.* **719**, 865–872 (2010).
43. von Zahn, U., Hunten, D. M. & Lehmacher, G. Helium in Jupiter's atmosphere: results from the Galileo probe helium interferometer experiment. *J. Geophys. Res.* **103**, 22815–22829 (1998).





**Extended Data Figure 1 | Validation of the calculation of gravitational harmonics with the CEPAM method.** The four panels provide a comparison of gravitational harmonics  $J_2$  to  $J_{10}$  calculated with various methods: CEPAM models with 241 radial layers (black points), CMS

models with 800 layers (grey points), CEPAM models with 1,041 layers (red crosses), and CMS calculations for the CEPAM models with 1,041 layers (blue circles).



**Extended Data Figure 2 | Constraint on the characteristic amplitude of deep differential rotation in Jupiter.** The red curves show the likelihood of models ( $y$  axis) in which to the differentially rotating outer region constrained by Juno's odd harmonics<sup>6</sup> we add a deeper cylindrical flow of amplitude  $v$  ( $x$  axis). The dashed red curve uses  $1\sigma$  error bars. The solid red curve considers an extended ensemble of possibilities for the outer flow<sup>6</sup> with solutions up to  $3\sigma$ . In both cases, the model favours  $v < 6 \text{ m s}^{-1}$ . The blue curve shows the same model but without the added outer layer. That model also favours low-amplitude winds but is found to be  $4 \times 10^4$  times less likely than the model including the differentially rotating outer region.

Extended Data Table 1 | Parameters used for the calculation of interior models

Parameter	Description	Type	Mean	$\sigma$ or $\Delta X$
$T_{1\text{bar}}$	1 bar temperature, from Voyager and Galileo measurements <sup>39,41</sup>	Gaussian	165K	4K
$Y_{\text{proto}}/(X_{\text{proto}} + Y_{\text{proto}})$	Protosolar helium mixing ratio obtained from solar models <sup>42</sup>	Gaussian	0.277	0.006
$Y_{\text{atm}}/(X_{\text{atm}} + Y_{\text{atm}})$	Helium mixing ratio in Jupiter's atmosphere as measured by the Galileo probe <sup>43</sup>	Gaussian	0.238	0.005
$P_{\text{He}}$	Characteristic pressure of the helium phase separation region <sup>29,32</sup>	Uniform	1.9 Mbar	1.1 Mbar
$\Delta T_{\text{He}}$	Temperature increase over the helium phase separation region <sup>12</sup>	Uniform	0	500 K
$L_{\text{dilcore}}$	Presence of the diluted core region	Binary	0/1	
$P_{\text{dilcore}}$	Pressure of the diluted core region	Uniform	21.5 Mbar	18.5 Mbar
$\Delta \log P_{\text{dilcore}}$	Smoothness of the diluted core transition	Uniform	0.0255	0.0245
$\Delta Z_{\text{dilcore}}$	Mass mixing ratio increase in the diluted core region	Uniform	0.2	0.2
$Z_{\text{ices}}^{(1)}$	Mass mixing ratio of ices in region (1)	Fitted		
$Z_{\text{rocks}}^{(1)}$	Mass mixing ratio of rocks in region (1)	Uniform	0.025	0.025
$\Delta Z_{\text{ices}}$	Jump in the mass mixing ratio of ices from region (1) to region (2)	Uniform	0.075	0.075
$\Delta Z_{\text{rocks}}$	Jump in the mass mixing ratio of rocks from region (1) to region (2)	Uniform	0.075	0.075
$M_{\text{core}}$	Mass of the central dense core	Fitted		

Data are from refs 12, 29, 32, 39, 41–43.



Extended Data Table 2 | Comparison of model gravitational harmonics

Model	Method	$J_2 \times 10^6$	$J_4 \times 10^6$	$J_6 \times 10^6$	$J_8 \times 10^6$	$J_{10} \times 10^6$	$J_{12} \times 10^6$
Polytrope	CEP	13988.65	-531.8675	30.06605	-1.98248	0.14772	-0.01201
	WH16	13988.51	-531.8281	30.11832	-2.13212	0.17407	-0.01568
	CEP-WH16	0.14	-0.0394	-0.05227	0.14964	-0.02635	0.00367
REOS1a	CEP	14696.72	-587.8227	34.22564	-2.29778	0.17296	-0.01413
	CMS	14690.66	-587.3989	34.26170	-2.46234	0.20218	-0.01821
	CEP-CMS	6.06	-0.4238	-0.03606	0.16456	-0.02922	0.00408
REOS1b	CEP	14702.78	-588.1331	34.24635	-2.29937	0.17309	-0.01414
	CMS	14696.72	-587.7090	34.28245	-2.46399	0.20232	-0.01822
	CEP-CMS	6.06	-0.4240	-0.03610	0.16462	-0.02923	0.00408
REOS1a-1b	CEP-CMS	-0.00	-0.1136	-0.05681	0.16621	-0.02936	0.00409
MH13	CEP	14695.97	-590.2377	34.46524	-2.31752	0.17465	-0.01428
	CMS	14690.96	-589.9033	34.51000	-2.48443	0.20422	-0.01841
	CEP-CMS	5.01	-0.3343	-0.04476	0.16691	-0.02957	0.00413

# Monolayer atomic crystal molecular superlattices

Chen Wang<sup>1</sup>, Qiyuan He<sup>2</sup>, Udayabagya Halim<sup>2</sup>, Yuanyue Liu<sup>3†</sup>, Enbo Zhu<sup>1</sup>, Zhaoyang Lin<sup>2</sup>, Hai Xiao<sup>3</sup>, Xidong Duan<sup>4</sup>, Ziyang Feng<sup>2</sup>, Rui Cheng<sup>1</sup>, Nathan O. Weiss<sup>1</sup>, Guojun Ye<sup>5</sup>, Yun-Chiao Huang<sup>1</sup>, Hao Wu<sup>1</sup>, Hung-Chieh Cheng<sup>1</sup>, Imran Shakir<sup>6</sup>, Lei Liao<sup>4</sup>, Xianhui Chen<sup>5</sup>, William A. Goddard III<sup>3</sup>, Yu Huang<sup>1,7</sup> & Xiangfeng Duan<sup>2,7</sup>

**Artificial superlattices, based on van der Waals heterostructures of two-dimensional atomic crystals such as graphene or molybdenum disulfide, offer technological opportunities beyond the reach of existing materials<sup>1–3</sup>. Typical strategies for creating such artificial superlattices rely on arduous layer-by-layer exfoliation and restacking, with limited yield and reproducibility<sup>4–8</sup>. The bottom-up approach of using chemical-vapour deposition produces high-quality heterostructures<sup>9–11</sup> but becomes increasingly difficult for high-order superlattices. The intercalation of selected two-dimensional atomic crystals with alkali metal ions offers an alternative way to superlattice structures<sup>12–14</sup>, but these usually have poor stability and seriously altered electronic properties. Here we report an electrochemical molecular intercalation approach to a new class of stable superlattices in which monolayer atomic crystals alternate with molecular layers. Using black phosphorus as a model system, we show that intercalation with cetyl-trimethylammonium bromide produces monolayer phosphorene molecular superlattices in which the interlayer distance is more than double that in black phosphorus, effectively isolating the phosphorene monolayers. Electrical transport studies of transistors fabricated from the monolayer phosphorene molecular superlattice show an on/off current ratio exceeding 10<sup>7</sup>, along with excellent mobility and superior stability. We further show that several different two-dimensional atomic crystals, such as molybdenum disulfide and tungsten diselenide, can be intercalated with quaternary ammonium molecules of varying sizes and symmetries to produce a broad class of superlattices with tailored molecular structures, interlayer distances, phase compositions, electronic and optical properties. These studies define a versatile material platform for fundamental studies and potential technological applications.**

We used black phosphorus (BP) as a model system to explore the molecular intercalation approach to produce monolayer phosphorene molecular superlattices (MPMS). BP has attracted considerable recent interest as a 2D layered semiconductor with tunable bandgap and respectable carrier mobility up to 1,000 cm<sup>2</sup> V<sup>−1</sup> s<sup>−1</sup> that could be used for next-generation electronics and optoelectronics<sup>15–17</sup>. But the difficulty in isolating and stabilizing monolayer phosphorene<sup>18,19</sup> has limited most studies to multi-layer BP flakes. The expected intrinsic properties of monolayer phosphorene, including large direct bandgap or high mobility, have been difficult to reach in typical exfoliated materials<sup>18,20</sup>. The intercalation of BP with organic molecules (such as cetyl-trimethylammonium bromide, CTAB) may effectively decouple the atomic layers, allowing access to monolayer phosphorene characteristics.

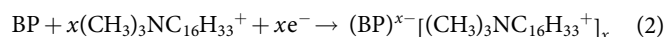
A home-designed electrochemical-optical measurement platform was used for *in situ* monitoring of the intercalation dynamics and the corresponding evolution of electronic and optical properties

(Fig. 1a, b). Briefly, a standard back-gated BP field-effect transistor (FET) was first fabricated on the SiO<sub>2</sub>/Si substrate and immersed in a polydimethylsiloxane reservoir filled with saturated CTAB solution in *N*-methyl-2-pyrrolidone, which is coupled with a platinum counter-electrode and an Ag/AgCl reference electrode, and placed under a Raman/photoluminescence microscope. During the intercalation process, the electrochemical current from the platinum counter-electrode, the drain current from the metal drain electrode and Raman/photoluminescence spectra can be simultaneously monitored *in situ*.

The electrochemical reaction consists of two half-reactions:



and



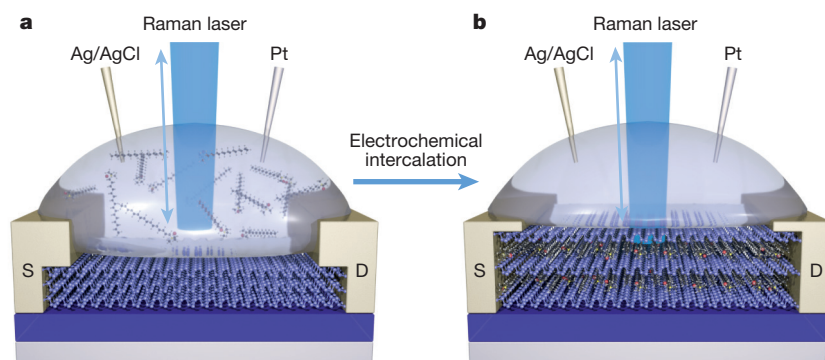
The electrochemical current was monitored *in situ* from the platinum counter-electrode (electrolyte gate) as the voltage was swept from 0 V to about 3 V. A careful analysis of the electrochemical gate current curve ( $I_{\text{eg}}$  in Fig. 2a) and its first derivative (Extended Data Fig. 1) shows apparent stepwise reactions, which may be attributed to the strong dependence of the bandgap and electronic properties of the BP on the layer number (see Methods for more detailed discussions). The stepwise reaction can be partitioned into six regions based on minimum points of the first derivative of the electrochemical current (Fig. 2a and Extended Data Fig. 1):

- 0.0–1.0 V: no obvious intercalation (because of the over-potential for the Br<sup>−</sup> sub-reaction; see more details in Methods);
- 1.0–1.4 V: major bulk intercalation occurring;
- 1.4–2.0 V: few-layer (four- to ten-layer) BP formation, less than 5 nm thick;
- 2.0–2.5 V: trilayer BP formation;
- 2.5–3.0 V: bilayer BP formation;
- beyond 3.0 V: monolayer phosphorene formation.

Figure 2b shows the corresponding source–drain current of the first scan (black) and the last scan (green) in a multi-scan intercalation process, demonstrating a relative small on/off ratio for the current in the starting BP before intercalation and strong gate modulation (on/off ratio of about 10<sup>5</sup>, limited by precise off-current measurement with the electrolyte gate) in the final fully intercalated materials.

*In situ* monitoring of the photoluminescence during the intercalation process reveals a gradual evolution from the absence of apparent photoluminescence emission in the visible to near-infrared regime in bulk BP to prominent emission at about 898 nm (1.38 eV), about 710 nm (1.75 eV) and eventually about 548 nm (2.26 eV) as the intercalation process progresses (Fig. 2c). These emission peaks correspond roughly

<sup>1</sup>Department of Materials Science and Engineering, University of California, Los Angeles, California 90095, USA. <sup>2</sup>Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, USA. <sup>3</sup>Materials and Process Simulation Center, California Institute of Technology, Pasadena, California 91125, USA. <sup>4</sup>State Key Laboratory for Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, School of Physics and Electronics, Hunan University, Changsha 410082, China. <sup>5</sup>Key Laboratory of Strongly Coupled Quantum Matter Physics, Hefei National Laboratory for Physical Science at Microscale and Department of Physics, University of Science and Technology of China, Hefei, Anhui 230026, China. <sup>6</sup>Sustainable Energy Technologies Centre, College of Engineering, King Saud University, Riyadh 11421, Kingdom of Saudi Arabia. <sup>7</sup>California Nanosystems Institute, University of California, Los Angeles, California 90095, USA. <sup>†</sup>Present addresses: Texas Materials Institute and Department of Mechanical Engineering, The University of Texas at Austin, Austin, Texas 78712, USA.



**Figure 1 | *In situ* electrochemical–optical measurement platform to monitor electrochemical intercalation process in real time.**

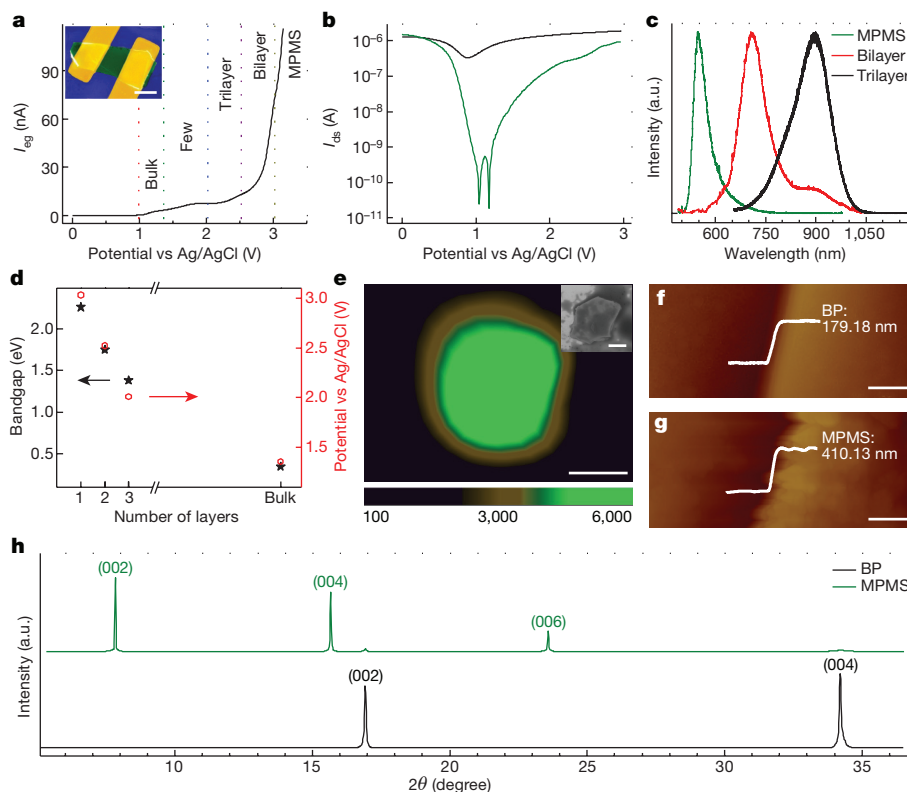
**a, b,** Schematics of the platform for the BP intercalation process from BP

(a) to MPMS (b). The system can allow simultaneous monitoring of the electrochemical current, source–drain current and photoluminescence/Raman spectra during the electrochemical intercalation process.

to near-band-edge emission from trilayer, bilayer and monolayer phosphorene. Interestingly, the bandgap and the onset intercalation potential show very similar relationships with the layer number (Fig. 2d), suggesting a close correlation between electrochemical potential and the measured bandgap. To the best of our knowledge, the photoluminescence emission at 548 nm (2.26 eV) indicates the highest optical bandgap observed in phosphorene or thin BP-based structures (values ranging from 1.45 eV to 1.84 eV were observed previously)<sup>18,20,21</sup>, indicating that we may have formed the true monolayer material. The observed optical bandgap of 2.26 eV is slightly larger than the theoretical limit of ideal phosphorene (about 2 eV)<sup>22</sup>. This difference may

be partly attributed to strain-induced bandgap expansion and orbital symmetry breaking caused by the insertion of CTAB and will be further discussed below<sup>20,23</sup>. Photoluminescence maps (at 550 nm) of the MPMS flake show highly uniform contrast (Fig. 2e), indicating the relative structure uniformity of the resulting MPMS.

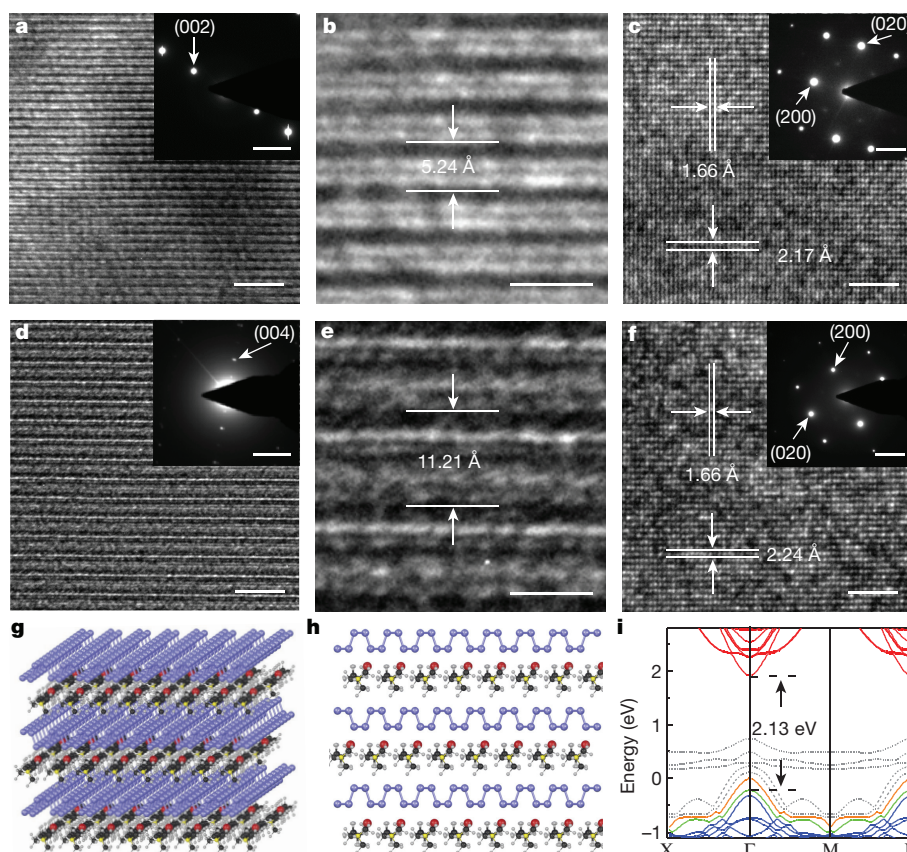
Atomic force microscope (AFM) studies show that, compared with pristine BP (Fig. 2f), the thickness of the fully intercalated MPMS (Fig. 2g) is increased by about 130%, which is consistent with X-ray diffraction (XRD) studies showing an 115% increase in the interlayer distance from 5.23 Å in BP to 11.27 Å in MPMS (Fig. 2h). Elemental analyses based on energy-dispersive X-ray (EDX) studies give an atomic ratio of



**Figure 2 | Structural and property evolution from BP to MPMS during the dynamic intercalation process.** **a,** Electrochemical gate current as a function of the applied electrochemical potential. Inset: false-colour scanning electron micrograph (SEM) of MPMS transistors. Scale bar: 5  $\mu\text{m}$ . **b,** Ionic gate transfer characteristics of the starting BP (black) and the final MPMS (green) at 0.01 V drain–source bias. **c,** Photoluminescence signals observed during different stages of intercalation. **d,** The relationship between bandgap or electrochemical potential and layer

number, showing a good correlation between electrochemical potential and the corresponding bandgap in the bulk, bilayer, trilayer and monolayer regimes. **e,** Photoluminescence mapping centred at 550.0 nm with similar intensity. Inset: the corresponding SEM image. Scale bars: 3  $\mu\text{m}$ . **f, g,** AFM images of a BP flake (**f**) and the resulting MPMS after CTAB intercalation (**g**). Scale bar: 300 nm. **h,** XRD pattern of BP and MPMS, verifying the interlayer distance expansion. a.u., arbitrary units.





**Figure 3 | TEM characterization of structure evolution from BP to MPMS.** **a, d**, Cross sectional TEM image comparison between BP (**a**) and MPMS (**d**). Scale bars: 3 nm. Insets are the corresponding electron diffraction pattern. Scale bars:  $2 \text{ nm}^{-1}$ . **b, e**, The corresponding high-resolution cross-sectional TEM images. Scale bars: 1 nm. **c, f**, Planar TEM images of BP (**c**) and MPMS (**f**) showing the lattice parameter expansion in the armchair direction and negligible change in zigzag direction. Scale

bars: 2 nm. Insets: electron diffraction of the corresponding TEM images; Scale bars:  $5 \text{ nm}^{-1}$ . **g, h**, Three-dimensional (**g**) and cross-sectional (**h**) views of the simulated atomic structure of MPMS. **i**, The simulated electronic structure of MPMS, demonstrating the enlarged bandgap of 2.13 eV in MPMS as determined by the transition between first valence-band maximum, VBM-1 (green), and conduction-band minimum, CBM (red).

P:N:Br of approximately 33.2:1.2:1.0 in MPMS (Extended Data Fig. 2), suggesting that Br is also intercalated into the final superlattice structure. Although free  $\text{Br}_2$  is produced and not intercalated into BP during the electrochemical intercalation process (see equations (1) and (2)), the  $\text{Br}_2$  produced during the electrochemical process may back-react with phosphorene-CTA<sup>+</sup> after releasing electrochemical potential; in this process,  $\text{Br}_2$  is reduced back to  $\text{Br}^-$  ions and intercalated into the final MPMS structure to form phosphorene/CTAB superlattices (see Methods for more details).

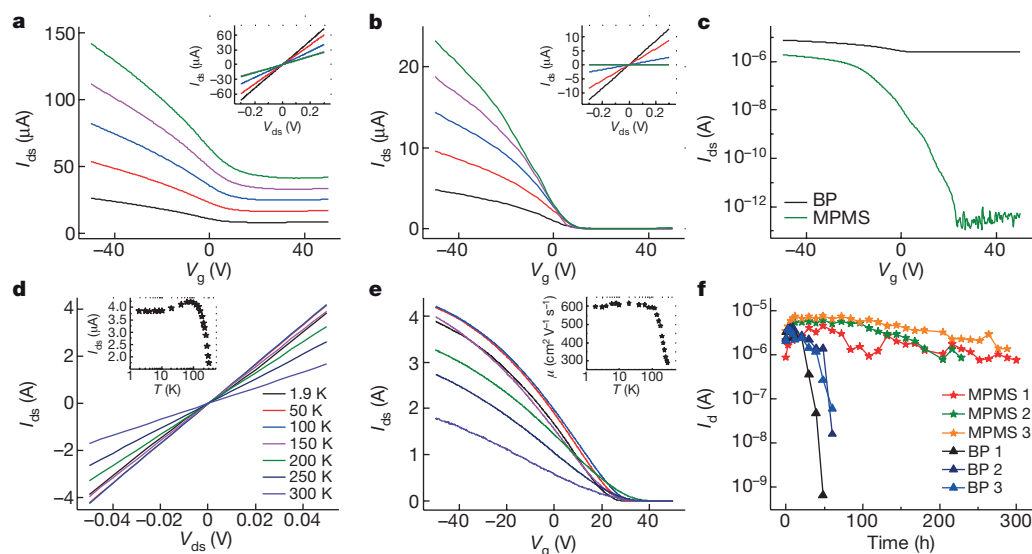
Studies with cross-sectional transmission electron microscopy (TEM) and electron diffraction (Fig. 3a, d) show the distinct differences in structure between the layered BP and the MPMS, with a clearly resolved interlayer distance expansion from 5.24 Å in BP (Fig. 3a, b) to 11.21 Å in MPMS (Fig. 3d, e), consistent with AFM and XRD studies (Fig. 2f–h). The increase of about 6 Å in the interlayer distance corresponds roughly to the end-to-end distance between the CTAB methyl–methyl substituents. The planar TEM studies reveal a lattice expansion of about 3% from 2.17 Å in pristine BP (Fig. 3c) to 2.24 Å in MPMS (Fig. 3f) in the armchair direction (200) and negligible change in the zigzag direction, which is further confirmed by the corresponding electron diffraction patterns (Fig. 3c, f, insets). This distinct expansion of about 3% in the armchair direction is also consistent with the photoluminescence (blueshift partly due to strain-induced bandgap expansion<sup>20</sup>) and the Raman spectroscopic studies (see Methods and Extended Data Fig. 3).

Density functional theory (DFT) calculations of the MPMS predict a relaxed structure with an interlayer distance of 11.41 Å, matching

well with the 11.27 Å value determined experimentally (Fig. 3g, h). Furthermore, the calculated MPMS structure also shows a 2.9% expansion in the armchair direction compared with that of BP calculated using the same method, consistent with the planar TEM observations and Raman spectroscopic studies (see Methods for more details). We attribute the strain to the repulsion between CTAB molecules, which leads to the expansion of the BP lattice, similar to the strain observed in alkali-metal intercalated graphite<sup>24</sup>. The electronic structure calculations show that the MPMS exhibits a bandgap of 2.13 eV (Fig. 3i), about 0.19 eV larger than that of monolayer phosphorene (1.94 eV) (see Methods and Extended Data Fig. 4), which is in agreement with the experimentally observed bandgap of 2.26 eV in MPMS.

We have also studied the electrical properties of the same back-gated BP and MPMS FETs before and after the intercalation process. In general, the output characteristics ( $I_{\text{ds}}-V_{\text{ds}}$ ) for both the BP and the MPMS devices show linear relationships (insets of Fig. 4a, b), suggesting no obvious contact barrier. The transfer characteristics of pristine BP show typical p-type behaviour with an on/off ratio <10 and a mobility value up to  $721 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  (Fig. 4a). The MPMS device retains p-type characteristics with a respectable mobility of  $328 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$  (Fig. 4b), which outperforms the best few-layer (<5 nm thick) BP devices and compares well with thin BP (5–15 nm thick) devices (see Extended Data Fig. 5 for a comparison of the electrical properties of MPMS devices with the recently reported few-layer and thin BP devices). Notably, the observed mobility in MPMS is also close to the theoretical limit predicted for monolayer phosphorene ( $250\text{--}400 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ )<sup>25</sup>. For the same BP before and after intercalation, the on/off ratio is greatly





**Figure 4 | Evolution of electrical properties from BP to MPMS and comparison of stability.** **a, b,** Back-gate transfer characteristics of pristine BP and MPMS as the source–drain bias is stepped from 0.1 V (black curve) to 0.5 V (green curve) bias. The insets show the corresponding output characteristics. **c,** Transfer characteristics at 0.01 V source–drain bias show

increased from  $<10$  in BP to  $>10^7$  in MPMS (Fig. 4c), at least one order of magnitude greater than previously achieved in phosphorene or thin BP devices<sup>15,19</sup>.

We have also explored the transport properties of MPMS FETs at various temperatures from 1.9 K to 300 K (Fig. 4d, e). With the decreasing temperature, the on-current more than doubled from 1.74  $\mu\text{A}$  at 300 K to 3.85  $\mu\text{A}$  at 1.9 K (inset of Fig. 4d), with the corresponding field-effect mobility increasing from 289  $\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$  at 300 K to 599  $\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$  at 1.9 K (inset of Fig. 4e). In the phonon-limited temperature range (100–300 K), the mobility best fits the expression  $\mu \propto T^{-\gamma}$ , with the exponent  $\gamma$  around 0.73 for MPMS. A power-law dependence with a positive exponent is indicative of a phonon-scattering mechanism and consistent with other studies of thin BP with band-like transport<sup>15,26</sup>.

With the sandwiching and encapsulation of monolayer phosphorene between molecular monolayers, the environmental stability is greatly increased. For example, comparing BP devices and MPMS devices with a similar on-current, the MPMS devices show little electrical degradation for as long as 300 h exposure in ambient conditions, whereas BP devices usually show serious degradation after 20–30 h exposure (Fig. 4f). The electrical stability of MPMS compares favourably with those of BN-encapsulated/passivated or  $\text{Al}_2\text{O}_3$ -passivated few-layer BP devices reported recently (Extended Data Table 1). The improved stability may be attributed to the special superlattice structure, in which the encapsulation of each phosphorene monolayer greatly slows the oxygen and water diffusion believed to be the main cause of BP degradation<sup>19,26–28</sup>. Together, the MPMS structure allows access to all key intrinsic characteristics of monolayer phosphorene, including high carrier mobility, high on/off current ratio, large optical bandgap and superior stability. It could open up new opportunities in phosphorene electronics and photonics. For example, by using lithography patterning to enable area selective intercalation, we have fabricated a lateral BP–MPMS heterojunction with diode-like rectification (see Methods and Extended Data Fig. 6), demonstrating a new pathway to functional phosphorene electronics and optoelectronics.

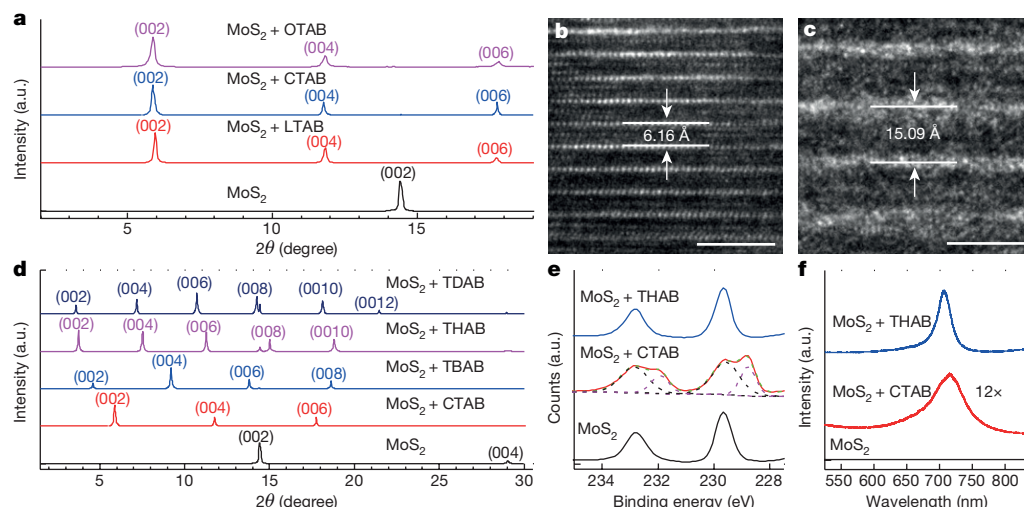
Our approach may be expanded to intercalate diverse 2D atomic crystals (2DACs) with various molecules to produce a broad range of monolayer atomic crystal molecular superlattices (MACMS). To this end, we have also intercalated  $\text{MoS}_2$  with quaternary ammonium molecules with variable carbon chain lengths (lauryl- or cetyl- or

an on/off ratio  $>10^7$  in MPMS versus  $<10$  in BP. **d, e,** Output and transfer characteristics of MPMS at various temperatures from 1.9 K (black) to 300 K (violet). Insets: on-current and mobility versus temperature. **f,** Comparison of electrical stability between three MPMS (stars) and three BP (triangles) devices with similar starting on-current.

octadecyl-trimethylammonium bromide: LTAB, CTAB or OTAB, respectively). The XRD (Fig. 5a) and HRTEM studies (Fig. 5b,c) of the resulted  $\text{MoS}_2$  molecular superlattices ( $\text{MoS}_2\text{MS}$ ) show that intercalation with these molecules produces a fairly similar interlayer distance expansion, from 6.14 Å in pristine  $\text{MoS}_2$  to 14.94 Å in the  $\text{MoS}_2/\text{LTAB}$  superlattice, 15.02 Å in  $\text{MoS}_2/\text{CTAB}$  and 14.99 Å in  $\text{MoS}_2/\text{OTAB}$ , suggesting that the molecule chains are oriented in parallel instead of vertical to the atomic layers, similar to those in the MPMS.

To tailor the interlayer distance, we have further explored tetrahedral-symmetry molecules with four equivalent substitutional groups (tetra-butyl, tera-heptyl or tetra-decyl)-ammonium bromide: TBAB, THAB or TDAB) for intercalating into  $\text{MoS}_2$ . The XRD studies clearly demonstrate a systematic expansion of interlayer distance from 6.14 Å in pristine  $\text{MoS}_2$ , to 15.02 Å in the  $\text{MoS}_2/\text{CTAB}$  superlattice, 19.15 Å in  $\text{MoS}_2/\text{TBAB}$ , 23.48 Å in  $\text{MoS}_2/\text{THAB}$  and 24.60 Å in  $\text{MoS}_2/\text{TDAB}$  (Fig. 5d). Besides the ability to tune interlayer distance, intercalation with different molecules could also produce materials with variable phase compositions. For example, the  $\text{MoS}_2/\text{CTAB}$  superlattice consists of mixed phases of metallic 1T- $\text{MoS}_2$  and semiconducting 2H- $\text{MoS}_2$ , whereas the  $\text{MoS}_2/\text{THAB}$  superlattice exhibits a pure semiconducting 2H- $\text{MoS}_2$  phase (Fig. 5e)<sup>29</sup>.

The intercalation with relatively large molecules to produce a superlattice structure can largely decouple the interlayer interaction and considerably modulate the electronic and optical properties of the 2DACs. For example, both the  $\text{MoS}_2/\text{CTAB}$  superlattice and  $\text{MoS}_2/\text{THAB}$  superlattice show a prominent photoluminescence emission that is more than two orders of magnitude stronger than that of pristine multi-layer  $\text{MoS}_2$  (Fig. 5f), suggesting a transformation from an indirect-bandgap semiconductor in multi-layer  $\text{MoS}_2$  to a direct-bandgap semiconductor in  $\text{MoS}_2/\text{CTAB}$  and the  $\text{MoS}_2/\text{THAB}$  superlattices. It is also noted that  $\text{MoS}_2/\text{THAB}$  exhibits a photoluminescence intensity 12 times stronger than that of the  $\text{MoS}_2/\text{CTAB}$  superlattice. The lower photoluminescence intensity in  $\text{MoS}_2/\text{CTAB}$  may be attributed to its mixed semiconducting and metallic phases, as compared with the pure semiconducting phase in  $\text{MoS}_2/\text{THAB}$ <sup>30</sup>. The resulting  $\text{MoS}_2/\text{THAB}$  superlattices, with nearly ideal monolayer characteristics (such as a direct bandgap), can be viewed as a new class of ‘bulk monolayer materials’ that are particularly attractive for efficient light-emitting devices.



**Figure 5 | Tunable structural and physical property of MACMS intercalated with different molecules.** **a**, XRD patterns of MoS<sub>2</sub> and MoS<sub>2</sub>MS intercalated with quaternary ammonium molecules with a variable carbon chain length (lauryl-, cetyl- or octadecyl-trimethylammonium bromide, that is, LTAB, CTAB or OTAB). **b**, **c**, Cross-sectional TEM image comparing pristine MoS<sub>2</sub> (**b**) and MoS<sub>2</sub>/CTAB superlattice (**c**). Scale bars: 2 nm. **d**, XRD patterns of MoS<sub>2</sub>

and MoS<sub>2</sub>MS intercalated with quaternary ammonium molecules with four equivalent substitutional groups of increasing chain length (tetra-butyl-, tetra-heptyl- or tetra-decyl-ammonium bromide; that is, TBAB, THAB or TDAB). **e**, XPS spectra of MoS<sub>2</sub>, MoS<sub>2</sub>/CTAB superlattice and MoS<sub>2</sub>/THAB superlattice. **f**, Photoluminescence signals observed from pristine MoS<sub>2</sub>, MoS<sub>2</sub>/CTAB superlattice and MoS<sub>2</sub>/THAB superlattice.

Lastly, we have shown that the same strategy can be used to intercalate a wide range of 2DACs, including 2D semiconductors (WSe<sub>2</sub>, In<sub>2</sub>Se<sub>3</sub>), 2D thermoelectric materials (SnSe), 2D multiferroic/piezo-electric materials (GeS), 2D semimetals (NbSe<sub>2</sub>), 2D superconductors (NbSe<sub>2</sub>), 2D charge-density-wave materials (NbSe<sub>2</sub>) and 2D topological insulators (Bi<sub>2</sub>Se<sub>3</sub>), with similar molecules to produce a broad class of MACMS with tailored molecular structure, interlayer distance, phase composition and electro-optical properties (Extended Data Fig. 7). Two kinds of superlattice structure were revealed in our studies of eight 2DACs intercalated with CTAB molecules: a type I structure with alternating monolayer 2DACs and monolayer molecules, as obtained in MPMS and In<sub>2</sub>Se<sub>3</sub>/CTAB, with interlayer distance expansion by about 6 Å; and a type II superlattice consisting of monolayer 2DACs and double layers of molecules, as observed in MoS<sub>2</sub>/CTAB, WSe<sub>2</sub>/CTAB, SnSe/CTAB, GeS/CTAB, NbSe<sub>2</sub>/CTAB or Bi<sub>2</sub>Se<sub>3</sub>/CTAB superlattice with interlayer gap expansion of about 9 Å. The origin of this difference will be an interesting topic for future studies.

Together, we have described a general electrochemical intercalation approach to a new class of MACMS consisting of alternating layers of monolayer 2DACs and molecular layers. By systematically varying the 2DACs and tailoring the molecular structures with varying sizes, symmetries, and substituent groups, a series of MACMS can be prepared with tailored interlayer distances, variable structure configurations and tunable electronic/optical properties. Furthermore, a wide range of functional molecules with different functional substituents/electronic structures or hybrid molecules with integrated functionalities, including magnetic molecules, photosensitive molecules, thermosensitive molecules and charge/energy-storage molecules, may be intercalated in selected 2DACs. Our studies thus provide an efficient route to such 2DACs–organic superlattices and define a versatile material platform for both fundamental physics studies and device applications. We hope that this will stimulate theoretical and experimental efforts to explore the broad choices of organic or inorganic intercalants that may produce MACMS with desirable electronic, optical and magnetic properties, thus offering a way to tailor and tame the properties of 2DACs.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 January 2017; accepted 17 January 2018.

- Novoselov, K. S., Mishchenko, A., Carvalho, A. & Neto, A. H. C. 2D materials and van der Waals heterostructures. *Science* **353**, aac9439 (2016).
- Liu, Y. *et al.* Van der Waals heterostructures and devices. *Nat. Rev. Mater.* **1**, 16042 (2016).
- Jariwala, D., Marks, T. J. & Hersam, M. C. Mixed-dimensional van der Waals heterostructures. *Nat. Mater.* **16**, 170–181 (2017).
- Haigh, S. J. *et al.* Cross-sectional imaging of individual layers and buried interfaces of graphene-based heterostructures and superlattices. *Nat. Mater.* **11**, 764–767 (2012).
- Yu, W. J. *et al.* Vertically stacked multi-heterostructures of layered materials for logic transistors and complementary inverters. *Nat. Mater.* **12**, 246–252 (2013).
- Cheng, R. *et al.* Electroluminescence and photocurrent generation from atomically sharp WSe<sub>2</sub>/MoS<sub>2</sub> heterojunction p–n diodes. *Nano Lett.* **14**, 5590–5597 (2014).
- Lee, C. H. *et al.* Atomically thin p–n junctions with van der Waals heterointerfaces. *Nat. Nanotech.* **9**, 676–681 (2014).
- Withers, F. *et al.* Light-emitting diodes by band-structure engineering in van der Waals heterostructures. *Nat. Mater.* **14**, 301–306 (2015).
- Gong, Y. *et al.* Vertical and in-plane heterostructures from WS<sub>2</sub>/MoS<sub>2</sub> monolayers. *Nat. Mater.* **13**, 1135–1142 (2014).
- Li, M.-Y. *et al.* Epitaxial growth of a monolayer WSe<sub>2</sub>–MoS<sub>2</sub> lateral p–n junction with an atomically sharp interface. *Science* **349**, 524–528 (2015).
- Duan, X. *et al.* Lateral epitaxial growth of two-dimensional layered semiconductor heterojunctions. *Nat. Nanotech.* **9**, 1024–1030 (2014).
- Bao, W. *et al.* Approaching the limits of transparency and conductivity in graphitic materials through lithium intercalation. *Nat. Commun.* **5**, 4224 (2014).
- Yu, Y. J. *et al.* Gate-tunable phase transitions in thin flakes of 1T-TaS<sub>2</sub>. *Nat. Nanotech.* **10**, 270–276 (2015).
- Xiong, F. *et al.* Li intercalation in MoS<sub>2</sub>: *in situ* observation of its dynamics and tuning optical and electrical properties. *Nano Lett.* **15**, 6777–6784 (2015).
- Li, L. K. *et al.* Black phosphorus field-effect transistors. *Nat. Nanotech.* **9**, 372–377 (2014).
- Perello, D. J., Chae, S. H., Song, S. & Lee, Y. H. High-performance *n*-type black phosphorus transistors with type control via thickness and contact-metal engineering. *Nat. Commun.* **6**, 7809 (2015).
- Yuan, H. *et al.* Polarization-sensitive broadband photodetector using a black phosphorus vertical p–n junction. *Nat. Nanotech.* **10**, 707–713 (2015).
- Pei, J. *et al.* Producing air-stable monolayers of phosphorene and their defect engineering. *Nat. Commun.* **7**, 10450 (2016).
- Ryder, C. R. *et al.* Covalent functionalization and passivation of exfoliated black phosphorus via aryl diazonium chemistry. *Nat. Chem.* **8**, 597–602 (2016).
- Liu, H. *et al.* Phosphorene: an unexplored 2D semiconductor with a high hole mobility. *ACS Nano* **8**, 4033–4041 (2014).
- Li, L. *et al.* Direct observation of the layer-dependent electronic structure in phosphorene. *Nat. Nanotech.* **12**, 21–25 (2016).

22. Castellanos-Gomez, A. Black phosphorus: narrow gap, wide applications. *J. Phys. Chem. Lett.* **6**, 4280–4291 (2015).
23. Guan, J., Song, W. S., Yang, L. & Tomanek, D. Strain-controlled fundamental gap and structure of bulk black phosphorus. *Phys. Rev. B* **94**, 045414 (2016).
24. Liu, Y., Merinov, B. V. & Goddard, W. A. Origin of low sodium capacity in graphite and generally weak substrate binding of Na and Mg among alkali and alkaline earth metals. *Proc. Natl Acad. Sci. USA* **113**, 3735–3739 (2016).
25. Rudenko, A. N., Brener, S. & Katsnelson, M. I. Intrinsic charge carrier mobility in single-layer black phosphorus. *Phys. Rev. Lett.* **116**, 246401 (2016).
26. Chen, X. L. *et al.* High-quality sandwiched black phosphorus heterostructure and its quantum oscillations. *Nat. Commun.* **6**, 7315 (2015).
27. Doganov, R. A. *et al.* Transport properties of pristine few-layer black phosphorus by van der Waals passivation in an inert atmosphere. *Nat. Commun.* **6**, 6647 (2015).
28. Avsar, A. *et al.* Air-stable transport in graphene-contacted, fully encapsulated ultrathin black phosphorus-based field-effect transistors. *ACS Nano* **9**, 4138–4145 (2015).
29. Acerce, M., Voiry, D. & Chhowalla, M. Metallic 1T phase MoS<sub>2</sub> nanosheets as supercapacitor electrode materials. *Nat. Nanotech.* **10**, 313–318 (2015).
30. Voiry, D. *et al.* Covalent functionalization of monolayered transition metal dichalcogenides by phase engineering. *Nat. Chem.* **7**, 45–49 (2015).

**Acknowledgements** The authors acknowledge the Electron Imaging Center for NanoMachines (EICN) at California NanoSystem Institute (CNSI) and Nanoelectronic Research Facility (NRF) at UCLA for technical support. Xiangfeng D. acknowledges support by National Science Foundation DMR1508144 (materials synthesis) and Office of Naval Research through grant number N00014-15-1-2368 (device fabrications). Y.H. acknowledges support by National Science Foundation EFRI-1433541. Y.L. was supported by a Resnick Prize Postdoctoral Fellowship at Caltech. L.L. acknowledges support through the 973 grant of MOST (No. 2013CBA01604). X.H.C. acknowledges support from

the National Natural Science Foundation of China (Grant No. 11534010). W.A.G. and Y.L. were also supported by DOE DE-SC0014607. W.A.G. acknowledges the Extreme Science and Engineering Discovery Environment (XSEDE) supported by National Science Foundation grant ACI-1053575. Y.L. acknowledges the computational resources sponsored by the DOE's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory, and the Texas Advanced Computing Center (TACC). I.S. thanks the Deanship of Scientific Research at King Saud University for its funding of this research through grant PEJP-17-01.

**Author Contributions** Xiangfeng D., Y.H. and C.W. co-designed the research. C.W. conducted device fabrication, electrical properties measurements and data analysis. C.W., Q.H. and U.H. conducted the intercalation experiments. C.W., U.H., Z.L. and Z.F. conducted structural and optical characterizations. Y.L., H.X. and W.A.G. contributed to the superlattice atomic and electronic structure calculations. E.Z. conducted the TEM studies. Q.H., Xidong D., Y.-C.H., H.W., H.-C.C., I.S. and L.L. contributed to the initial measurement system set-up, preparation of 2D materials and data analysis. R.C. contributed to the initial BP property characterization. N.O.W. contributed to the schematic drawing. G.J.Y. and X.H.C. prepared the initial BP material. Y.H. and Xiangfeng D. supervised the research. Xiangfeng D. and C.W. co-wrote the manuscript. All authors discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Xiangfeng D. ([xduan@chem.ucla.edu](mailto:xduan@chem.ucla.edu)), Y.H. ([yhuang@seas.ucla.edu](mailto:yhuang@seas.ucla.edu)) or L.L. ([liaolei@hnu.edu.cn](mailto:liaolei@hnu.edu.cn)).

**Reviewer Information** *Nature* thanks N. Guisinger, K. Loh and Q. Xiong for their contribution to the peer review of this work.



## METHODS

**2DAC device fabrication.** BP was synthesized under a constant pressure of 10 kbar by heating red phosphorus (99.999%) to 1,000 °C and slowly cooling to 600 °C at a cooling rate of 100 °C per hour. Other 2DACs were purchased from HQ Graphene. A standard mechanical exfoliation method was used to isolate thin BP on 300 nm SiO<sub>2</sub>/Si substrate in a nitrogen-filled glove-box for single-flake BP device study. The single thin BP transistors were fabricated using electron-beam lithography followed by electron-beam deposition of 10/80 nm Cr/Au metal thin films. The metal contact was patterned on top of BP edges and wrapped around BP edges. Other 2DACs flakes were isolated on the 100 nm Au/50 nm Ti/300 nm SiO<sub>2</sub>/Si substrate by standard mechanical exfoliation methods for electrochemical molecular intercalation.

**Structural characterizations.** SEM characterizations were performed by FEI Nova Nano 230 at a voltage of 15 kV. TEM cross-sectional samples of BP and MoS<sub>2</sub> were made by focused ion beam cutting from the thin BP and MoS<sub>2</sub>. TEM cross-sectional samples of MPMS and MoS<sub>2</sub>/CTAB superlattice were made through the sonication of thick CTAB-intercalated BP or MoS<sub>2</sub> flakes to make MPMS or MoS<sub>2</sub>/CTAB superlattice nanoprisms, which were then dipped onto TEM grids for imaging. Planar TEM samples were made by directly transferring the mechanically exfoliated thin BP to TEM grids that were directly used for intercalation and subsequent imaging. TEM characterizations were performed by FEI Titan at 300 kV accelerating voltage. XRD samples were prepared by transferring mechanically exfoliated thin 2DACs to 300 nm SiO<sub>2</sub>/Si substrate with a 100 nm/50 nm Ti/Au film on top, then characterized by PANalytical XPert Pro power X-ray diffractometer with 45 kV voltage, 40 mA emission current and 1/4° beam slit. The micro-Raman and micro-photoluminescence studies were conducted using a Horiba LabRAM HR Evolution confocal Raman system with Ar ion laser (488 nm and 633 nm) excitation. XPS spectra were collected from an Axis Ultra system equipped with monochromatic Al source under  $1 \times 10^{-8}$  torr vacuum.

**Intercalation chemistry of BP.** During the electrochemical intercalation process, the electrons are transferred from the highest occupied molecular orbital (HOMO) level of the reducing agent on anode to the lowest unoccupied molecular orbital (LUMO) of the BP on the cathode. On the anode, two bromide ions lose an electron to form Br<sub>2</sub>, requiring about 1 V of electrochemical potential. On the cathode side, electrons are pumped into the BP conduction band. The additional electrons cause BP to be negatively charged. To stabilize the additional electrons, (CH<sub>3</sub>)<sub>3</sub>NC<sub>16</sub>H<sub>33</sub><sup>+</sup> molecules insert themselves between the BP layers, acting as counter-ions.

The start of the major reaction observed at about 1 V can be explained by the difference of around 0.35 eV between the HOMO/valence band and LUMO/conduction band of BP and the Br<sup>−</sup> sub-reaction electrochemical potential of about 1 eV. The first insertion of (CH<sub>3</sub>)<sub>3</sub>NC<sub>16</sub>H<sub>33</sub><sup>+</sup> does not require high applied potential, owing to the low bandgap (approximately 0.35 eV) of the bulk BP. As the intercalation continues, the HOMO/LUMO gap of BP/substrate increases, making it harder for the next reaction to occur. The intercalation decouples the neighbouring BP layers and reduces the effective layer thickness, leading to a larger bandgap and a blueshift of the photoluminescence peak wavelength. A final peak was recorded at about 2.26 eV for MPMS, higher than previously observed optical gaps<sup>18,20,21</sup>.

During the reaction, it was clear that Br<sub>2</sub> was produced at the anode, as indicated by the emergence of the darker yellow/red Br<sub>2</sub> solution. However, after the electrochemical potential was withdrawn, the high activity of Br<sub>2</sub> means that it will back-react with phosphorene<sup>−</sup> CTA<sup>+</sup> layers quickly to form the final phosphorene/CTAB layered superlattice structure, which is consistent with the TEM EDX analysis and the structure simulated by DFT. The concentration change of Br<sub>2</sub> can be visually observed by the colour of the solution, from a transparent solution before reaction, to gradually darker yellow/red during the reaction, to weaker yellow after voltage withdrawal.

**Stepwise reaction mechanism.** Because the bandgap of BP is strongly dependent on the layer number, a stepwise reaction mechanism is proposed, mainly based on the characteristic stepwise electrochemical current curve<sup>31</sup> and the corresponding photoluminescence evolution from bulk, to few-layer, trilayer, bilayer and monolayer phosphorene as described in the main text. The CTA<sup>+</sup> with one positive charge and relative large molecular size was expected to intercalate in stepwise fashion, based on modelling study<sup>32</sup>.

For the intercalation experiments, the intercalation scan times, scan step, step duration and maximum voltage (stop voltage) can affect the final structure considerably. Owing to the diffusion-limited intercalation progress, the fully intercalated structure may not form within one fast scan. To better control the intercalation process, a multi-scan method was used in this study. In addition, too large a scan step may lead to sample cracking at a higher voltage. Too short a step duration time can cause incomplete intercalation, limited by the supply of CTAB molecules or by CTAB diffusion inside BP. Incorrect stop voltage will lead to strongly mixed phase, especially in the intermediate states. A stop voltage above 3 V for a long

time will result in sample cracking<sup>33</sup>. Multiple scans with different parameters may be used to tune the structure slowly from mixed phases to a relative pure phase by gradually increasing the intercalation degree. The exact intercalation parameters vary with sample size, thickness, shape (sharp or rough edges) and expected structure.

**Raman spectroscopic analysis of MPMS.** Raman studies show that all three characteristic BP Raman modes remain in MPMS after intercalation, yet with considerable intensity weakening (by a factor of about 40) (Extended Data Fig. 3a). A close analysis reveals considerable peak broadening and apparent shifts in peak position. The A<sub>g</sub><sup>1</sup> mode is slightly redshifted from BP 360.93 cm<sup>−1</sup> to MPMS 359.77 cm<sup>−1</sup> (Extended Data Fig. 3b), whereas the B<sub>2g</sub> and A<sub>g</sub><sup>2</sup> modes are blueshifted from 437.98 cm<sup>−1</sup> in BP to 438.21 cm<sup>−1</sup> in MPMS (Extended Data Fig. 3c) and 465.45 cm<sup>−1</sup> in BP to 465.96 cm<sup>−1</sup> in MPMS (Fig. 3d), respectively. As depicted in the atomic motions of three lattice vibrational modes (see the insets of Extended Data Fig. 3b–d), the armchair-direction strain of about 3% observed in the TEM analysis<sup>34</sup> will contribute positively to projected components of A<sub>g</sub><sup>1</sup> but negatively to the projected motion of A<sub>g</sub><sup>2</sup>. This leads to the redshift for A<sub>g</sub><sup>1</sup> and the blueshift for A<sub>g</sub><sup>2</sup>. Although atomic motions associated with B<sub>2g</sub> occur mostly along the zigzag direction, the armchair expansion will harden the zigzag-direction atomic motions indirectly, resulting in a very small blueshift of B<sub>2g</sub>. Therefore, the energy spacing between A<sub>g</sub><sup>1</sup> and B<sub>2g</sub> or between A<sub>g</sub><sup>1</sup> and A<sub>g</sub><sup>2</sup> modes increases under armchair stretching<sup>34</sup>.

**MPMS structure simulation methods.** According to DFT calculations using the Vienna Ab initio Simulation Package (VASP) with projector augmented-wave pseudopotentials, the lattice parameter in the armchair direction changed from 4.62 Å for BP to 4.75 Å for MPMS, representing a 2.9% increment. For the zigzag direction, the change is negligible on comparing 3.30 Å of BP with 3.28 Å of MPMS. For MPMS, we calculate a bandgap of 1.14 eV by using the Perdew–Burke–Ernzerhof (PBE) functional, compared with 0.90 eV for monolayer phosphorene. PBE underestimates bandgaps but indicates a bandgap increment of 0.24 eV for MPMS compared with monolayer phosphorene.

For an accurate evaluation of the electronic structure, we used the B3PW91 functional as implemented in the CRYSTAL14 package<sup>35,36</sup>, and all-electron 6-31G(d) basis sets of double- $\zeta$  quality were used for H, C, N, O, P and Br. For B3PW91 functional, an extra-large grid, consisting of 75 radial points and 974 angular points, was used for accurate integration, and the reciprocal space was sampled by the  $\Gamma$ -centred Monkhorst–Pack scheme with a  $7 \times 1 \times 3$  grid.

From Extended Data Fig. 4, there are additional bands within the bandgap of MPMS in Extended Data Fig. 4b that are marked by a grey dotted line, which are mainly composed of bromine atomic *p* orbitals, whereas the orange VBM-0 bands are mostly dominated by phosphorus orbitals. Our calculations of the frequency-dependent dielectric matrix (at the PBE level) indicate that in the isolated monolayer BP, the VBM-1 (Extended Data Fig. 4e) band and VBM-0 (Extended Data Fig. 4d) have symmetrically equivalent charge-density distributions, and thus they are close to each other in energy and both contribute to the first absorption peak (Extended Data Fig. 4a), as they both have considerable overlap with the CBM band state (Extended Data Fig. 4c). However, for MPMS, the VBM-1 (Extended Data Fig. 4h) has the largest overlap with CBM (Extended Data Fig. 4f) and high possibility of contribution to the transition from VBM-1 to CBM. For VBM-0 (Extended Data Fig. 4g) and CBM (Extended Data Fig. 4f), owing to overlap between states being small, the transition between them is very limited. The optical transition bandgap of MPMS was therefore determined by the transition from VBM-1 (Extended Data Fig. 4h, green band in Extended Data Fig. 4b) to CBM (Extended Data Fig. 4f, red band in Extended Data Fig. 4b), which matched well with our experimental observation of an enlarged bandgap in MPMS.

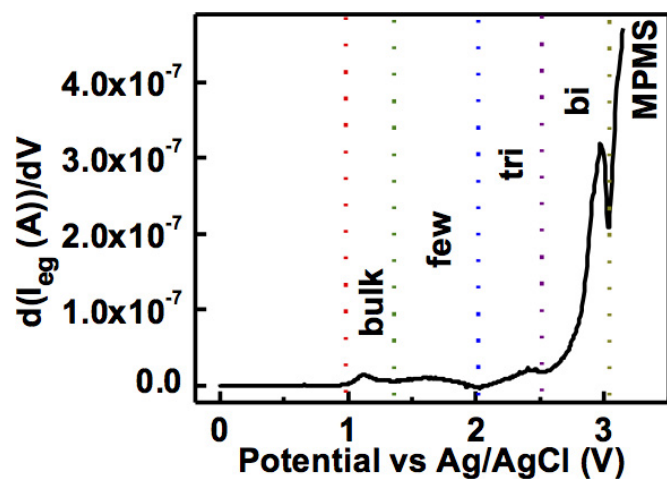
**Lateral BP-MPMS heterojunctions.** Lateral integration of 2D materials by using either the same material with different dopants<sup>37,38</sup> or two different materials with opposite doping type<sup>39</sup> is an exciting topic for creating functional nanodevices. However, lateral integration of phosphorene-based structure has not been sufficiently studied, owing to the difficulty in controllable and selective doping and the complexity in fabricating such devices. Inspired by superior electrical properties of MPMS, Fermi-level mismatch between BP and MPMS, and the high controllability of the intercalation process, we fabricated lateral BP-MPMS heterojunctions by partial interaction of a BP flake. Because the insertion of CTAB into BP occurs through the edges, partial intercalation can be achieved by selectively opening an intercalation window on a PMMA-covered BP and controlling the diffusion-limited intercalation time. The exposed edges underwent electrochemical reactions, forming a lateral junction between the intercalated MPMS and the passivated BP. Photoluminescence spectra mapping of a typical lateral BP-MPMS heterostructure show a clear photoluminescence signal in the MPMS region and the absence of the photoluminescence signal in the BP part (Extended Data Fig. 6a). Similarly, corresponding Raman spectra mapping centred at 438 cm<sup>−1</sup> showed a considerably



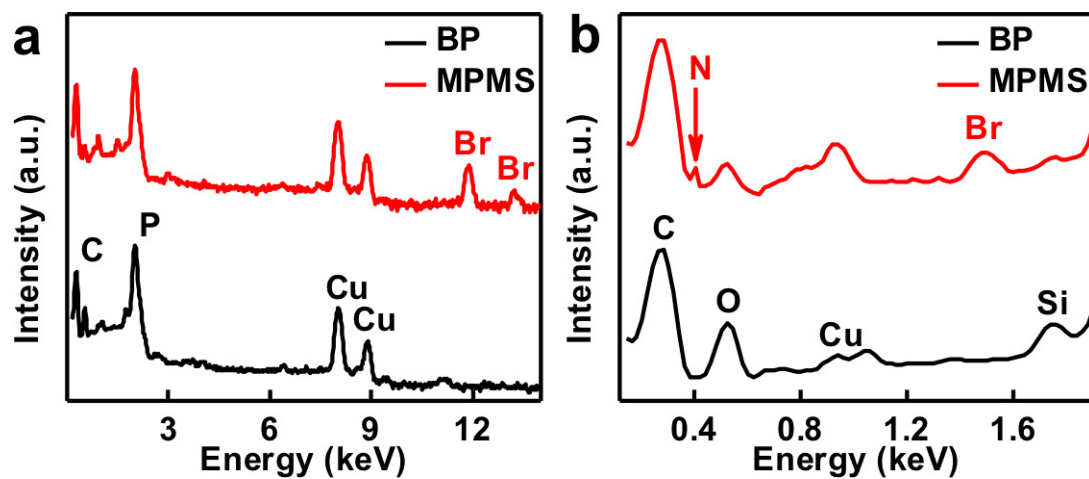
stronger signal in the BP region than that in the MPMS region (Extended Data Fig. 6b). After the standard electron beam lithography and metal deposition process to contact the BP region and MPMS region separately, we obtained a BP–MPMS heterojunction device (Extended Data Fig. 6c and inset of Extended Data Fig. 6f), which is schematically illustrated in the inset of Extended Data Fig. 6d. Considering the large bandgap difference between monolayer phosphorene (MPMS) and BP, a diode-like rectification was expected from the band diagram (Extended Data Fig. 6e) and was indeed observed (Extended Data Fig. 6f)<sup>40</sup>. This demonstration of a unique lateral BP–MPMS heterojunction diode represents an essential step towards functional phosphorene electronics and optoelectronics.

**Data availability.** The data that support the findings of this study are available from the corresponding author on reasonable request.

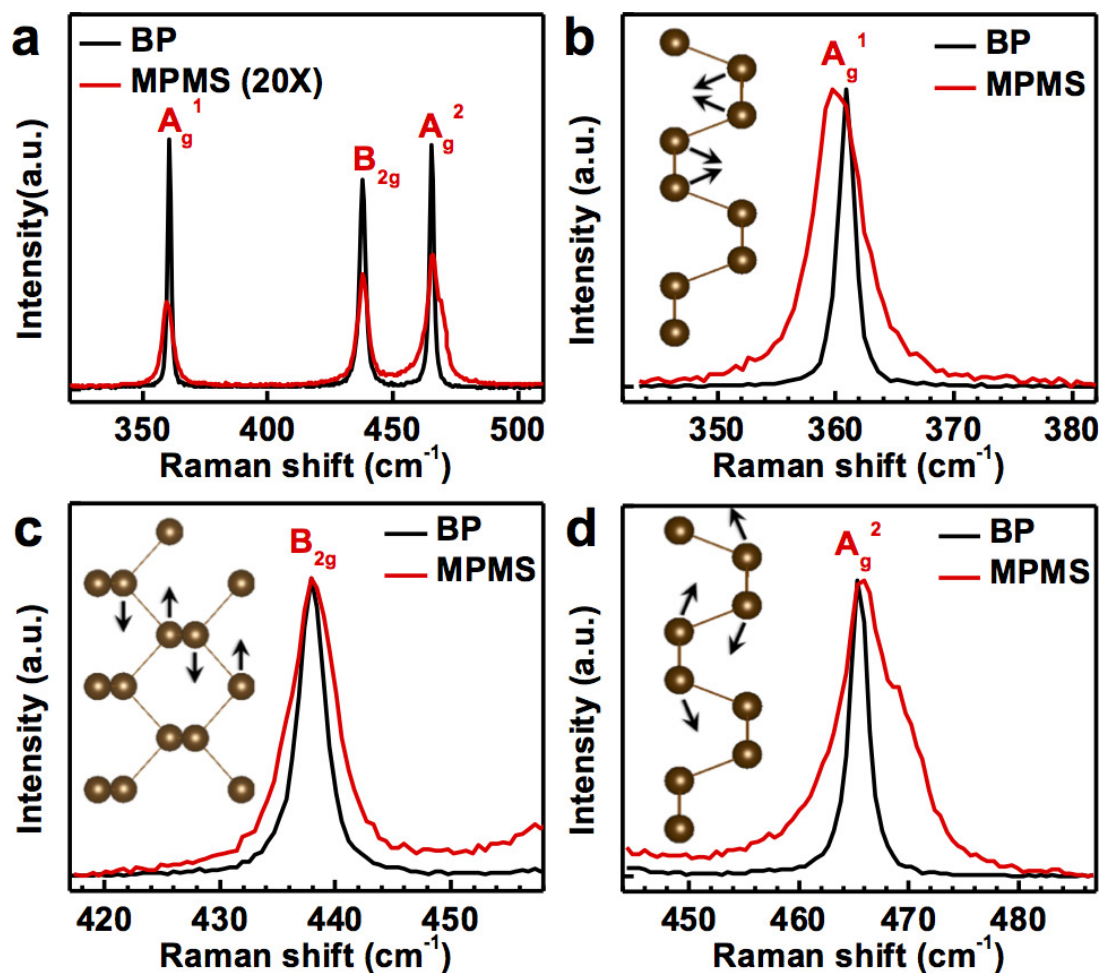
31. Sole, C., Drewett, N. E. & Hardwick, L. J. *In situ* Raman study of lithium-ion intercalation into microcrystalline graphite. *Faraday Discuss.* **172**, 223–237 (2014).
32. Hembram, K. P. S. S. *et al.* A comparative first-principles study of the lithiation, sodiation, and magnesiation of black phosphorus for Li-, Na-, and Mg-ion batteries. *Phys. Chem. Chem. Phys.* **18**, 21391–21397 (2016).
33. Hembram, K. P. S. S. *et al.* Unraveling the atomistic sodiation mechanism of black phosphorus for sodium ion batteries by first-principles calculations. *J. Phys. Chem. C* **119**, 15041–15046 (2015).
34. Fei, R. & Yang, L. Lattice vibrational modes and Raman scattering spectra of strained phosphorene. *Appl. Phys. Lett.* **105**, 083120 (2014).
35. Xiao, H., Tahir-Kheli, J. & Goddard, W. A. Accurate band gaps for semiconductors from density functional theory. *J. Phys. Chem. Lett.* **2**, 212–217 (2011).
36. Crowley, J. M., Tahir-Kheli, J. & Goddard, W. A. Resolution of the band gap prediction problem for materials design. *J. Phys. Chem. Lett.* **7**, 1198–1203 (2016).
37. Cai, J. M. *et al.* Graphene nanoribbon heterojunctions. *Nat. Nanotech.* **9**, 896–900 (2014).
38. Yu, W. J., Liao, L., Chae, S. H., Lee, Y. H. & Duan, X. F. Toward tunable band gap and tunable Dirac point in bilayer graphene with molecular doping. *Nano Lett.* **11**, 4759–4763 (2011).
39. Levendorf, M. P. *et al.* Graphene and boron nitride lateral heterostructures for atomically thin circuitry. *Nature* **488**, 627–632 (2012).
40. Cai, Y., Zhang, G. & Zhang, Y.-W. Layer-dependent band alignment and work function of few-layer phosphorene. *Sci. Rep.* **4**, 6677 (2014).
41. Yang, Z. B. *et al.* Field-effect transistors based on amorphous black phosphorus ultrathin films by pulsed laser deposition. *Adv. Mater.* **27**, 3748–3754 (2015).
42. Xia, F., Wang, H. & Jia, Y. Rediscovering black phosphorus as an anisotropic layered material for optoelectronics and electronics. *Nat. Commun.* **5**, 4458 (2014).
43. Kamalakar, M. V., Madhushankar, B. N., Dankert, A. & Dash, S. P. Low Schottky barrier black phosphorus field-effect devices with ferromagnetic tunnel contacts. *Small* **11**, 2209–2216 (2015).
44. Miao, J. S., Zhang, S. M., Cai, L., Scherr, M. & Wang, C. Ultrashort channel length black phosphorus field-effect transistors. *ACS Nano* **9**, 9236–9243 (2015).
45. Youngblood, N., Chen, C., Koester, S. J. & Li, M. Waveguide-integrated black phosphorus photodetector with high responsivity and low dark current. *Nat. Photon.* **9**, 247–252 (2015).
46. Na, J. *et al.* Few-layer black phosphorus field-effect transistors with reduced current fluctuation. *ACS Nano* **8**, 11753–11762 (2014).
47. Wood, J. D. *et al.* Effective passivation of exfoliated black phosphorus transistors against ambient degradation. *Nano Lett.* **14**, 6964–6970 (2014).
48. Buscema, M. *et al.* Fast and broadband photoresponse of few-layer black phosphorus field-effect transistors. *Nano Lett.* **14**, 3347–3352 (2014).
49. Kim, J. S. *et al.* Dual gate black phosphorus field effect transistors on glass for nor logic and organic light emitting diode switching. *Nano Lett.* **15**, 5778–5783 (2015).
50. Hong, T. *et al.* Polarized photocurrent response in black phosphorus field-effect transistors. *Nanoscale* **6**, 8978–8983 (2014).
51. Koenig, S. P., Doganov, R. A., Schmidt, H., Neto, A. H. C. & Ozyilmaz, B. Electric field effect in ultrathin black phosphorus. *Appl. Phys. Lett.* **104**, 103106 (2014).
52. Zhu, W. N. *et al.* Flexible black phosphorus ambipolar transistors, circuits and AM demodulator. *Nano Lett.* **15**, 1883–1890 (2015).
53. Wang, H. *et al.* Black phosphorus radio-frequency transistors. *Nano Lett.* **14**, 6424–6429 (2014).
54. Kah-Wee, A., Zhi-Peng, L. & Juntao, Z. Next generation field-effect transistors based on 2D black phosphorus crystal. *2015 IEEE Int. Conf. on Digital Signal Processing*, 1223–1226 (2015).
55. Viti, L. *et al.* Efficient terahertz detection in black-phosphorus nano-transistors with selective and controllable plasma-wave, bolometric and thermoelectric response. *Sci. Rep.* **6**, 20474 (2016).
56. Wan, B. S. *et al.* Enhanced stability of black phosphorus field-effect transistors with SiO<sub>2</sub> passivation. *Nanotechnology* **26**, 435702 (2015).



**Extended Data Figure 1 | Stepwise reaction mechanism and its partition map.** First derivative of the electrochemical gate current in Fig. 2a. By analysing the original current curve and local minimum of the first derivative, the stepwise reaction can be clearly identified: that is, no major intercalation for 0–1.0 V (over-potential for  $\text{Br}^-$  sub-reaction), 1.0–1.4 V for major bulk intercalation, 1.4–2.0 V for few-layer BP formation, 2.0–2.5 V for trilayer BP formation, 2.5–3.0 V for bilayer BP formation and beyond 3.0 V for MPMS formation, which is also consistent with bandgap evolution from bulk, few, trilayer and bilayer to monolayer phosphorene.



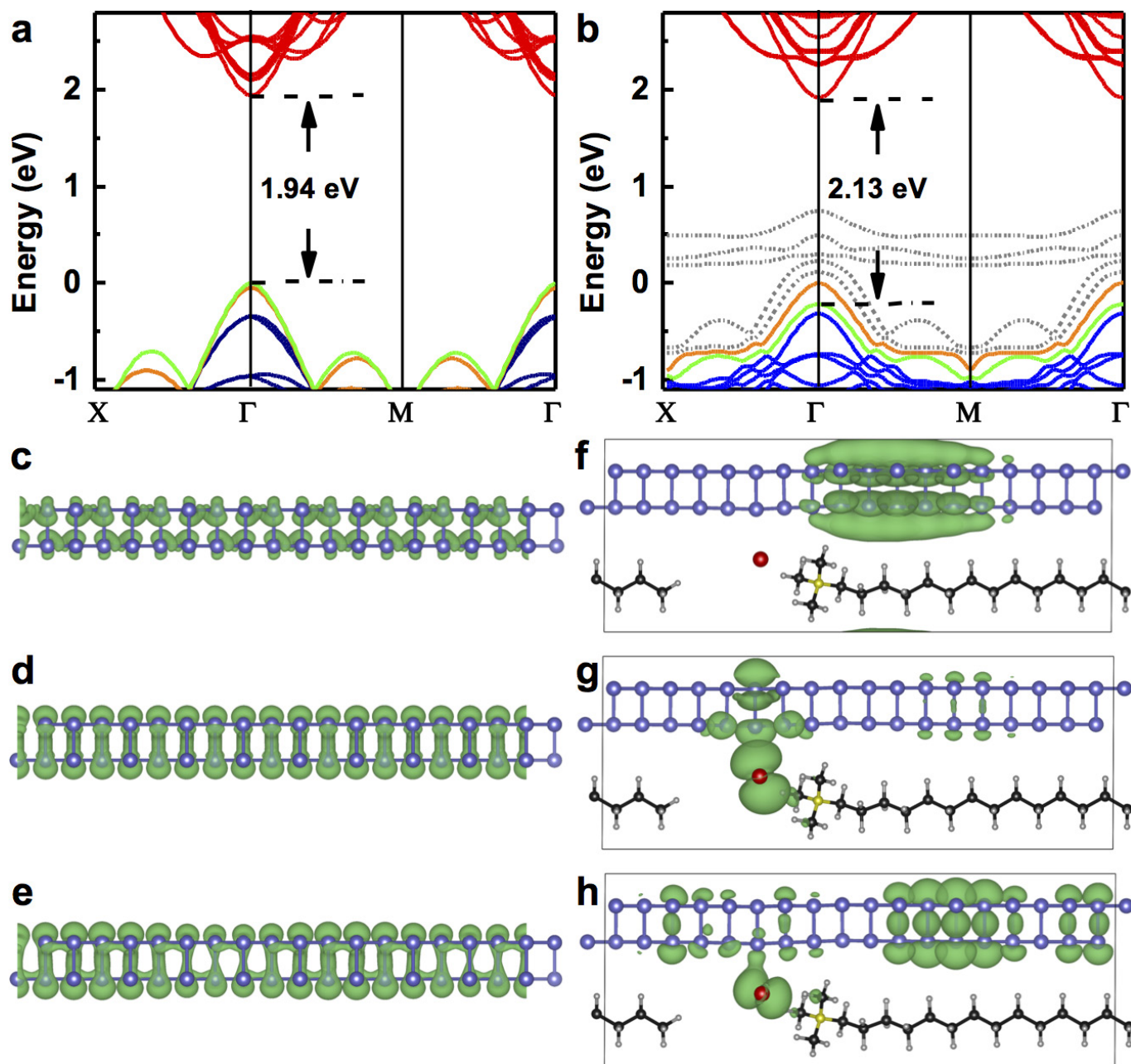
**Extended Data Figure 2 | TEM EDX spectra of BP and MPMS. a, b,** Spectra of BP and MPMS, showing the existence of Br and N after intercalation. Three average spectra gave an atomic ratio of P:N:Br as 33.2:1.2:1.0.



**Extended Data Figure 3 | Raman spectra characterization of BP and MPMS.** **a**, Raman spectra to compare the relative peak intensity and full-width at half-maximum evolution from pristine BP (black) to MPMS (red). The MPMS spectrum is multiplied by 20 for easy comparison.

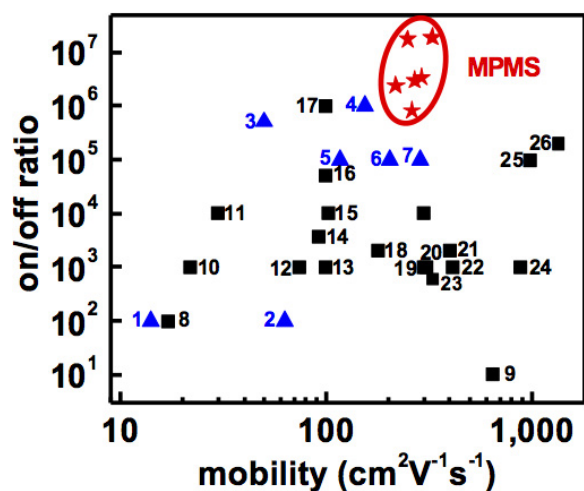
**b–d**,  $A_g^1$ ,  $B_{2g}$  and  $A_g^2$  mode comparison between pristine BP and MPMS to show redshift, blueshift and blueshift after MPMS formation, respectively. Insets: schematic illustration of atomic motion of each vibration modes.



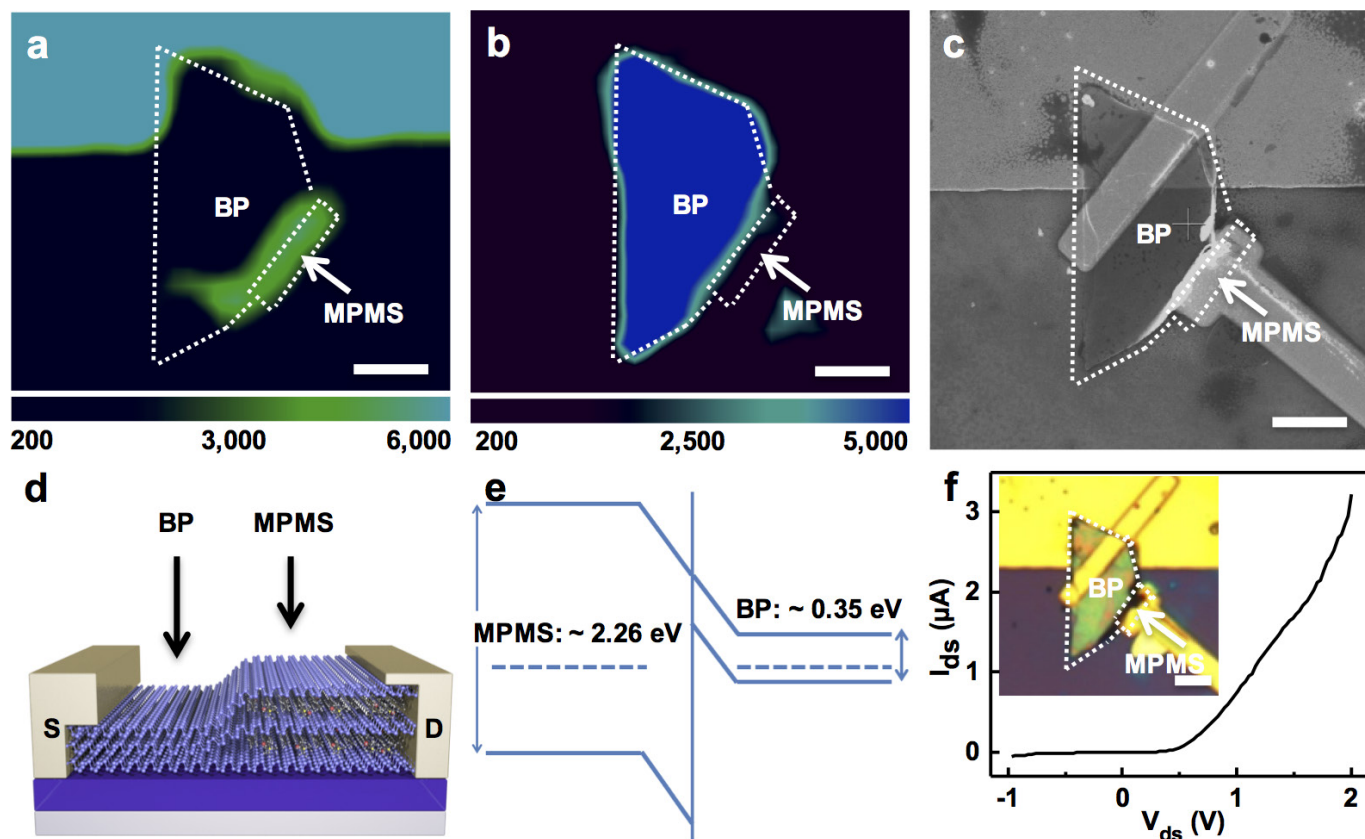


**Extended Data Figure 4 | The calculated electronic band structure evolution from BP to MPMS.** **a, b,** Electronic structure of monolayer phosphorene (**a**) and MPMS (**b**), demonstrating the enlarged bandgap from 1.94 eV in monolayer phosphorene to 2.13 eV in MPMS, as determined by the transition from VBM-1 (green) and CBM (red). The newly introduced bands of MPMS marked as grey dotted lines are mainly from bromine atomic *p* orbitals. The orange VBM-0 band is mainly (about 90%) from phosphorus, but those orbitals contribute little to the optical

transition, owing to very small overlap with the CBM. **c, d, e,** Monolayer phosphorene charge-density distribution of CBM (red in **a**), VBM-0 (orange in **a**) and VBM-1 (green in **a**), showing the transition bandgap determined by CBM and VBM-0/VBM-1 (very close in energy). **f, g, h,** MPMS charge-density distribution of CBM (red in **b**), VBM-0 (orange in **b**) and VBM-1 (green in **b**), showing the transition bandgap determined by VBM-1 and CBM due to large overlap of charge density.



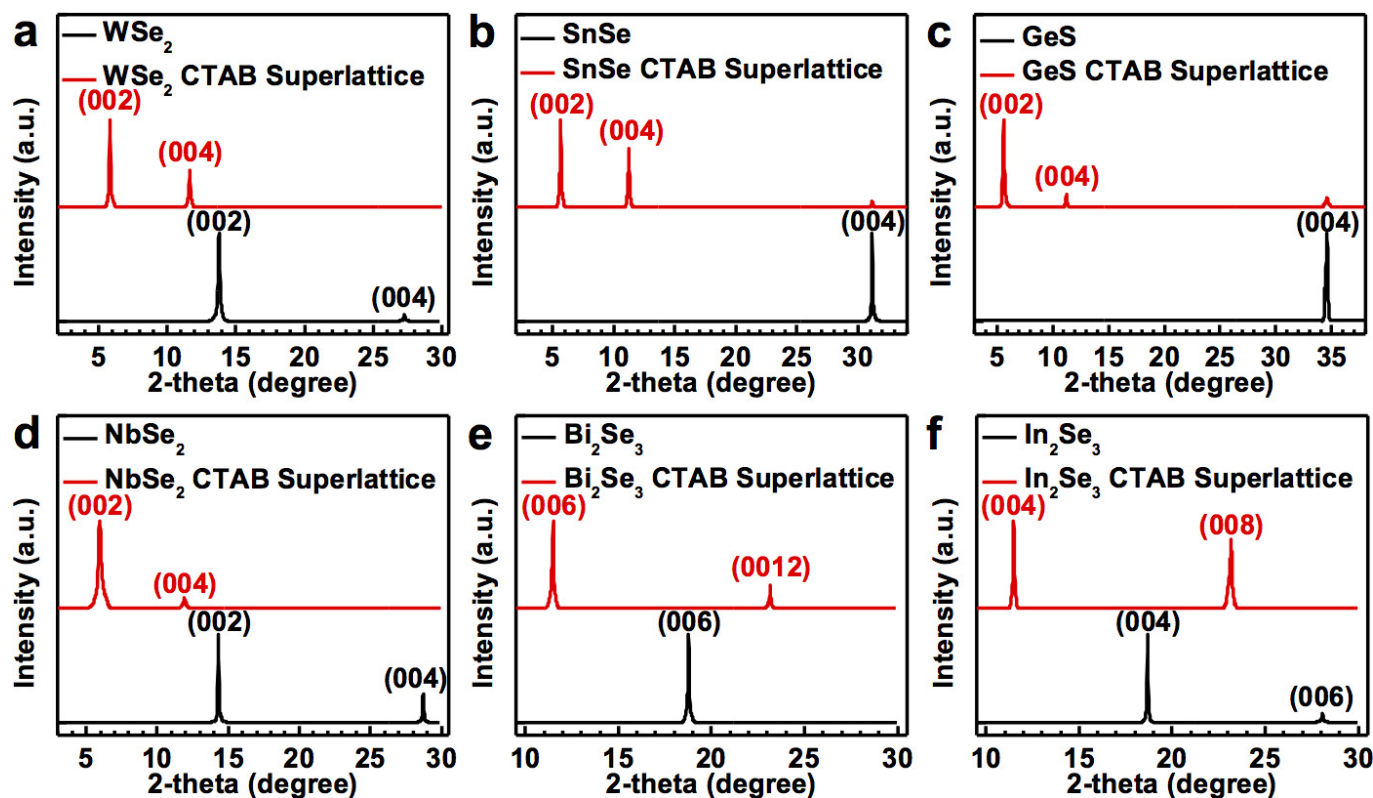
**Extended Data Figure 5 | The on/off ratio and mobility of the MPMS devices and the recently reported few-layer and thin BP devices.** Six MPMS devices (red star) show an average mobility of  $270 \text{ cm}^2 \text{V}^{-1} \text{s}^{-1}$  and averaged on/off ratio of  $8.6 \times 10^6$ . For comparison, we list recent studies of few-layer BP (less than 5 nm, marked as the blue triangle) and thin BP (5 nm to 15 nm, marked as the black square) devices. The MPMS devices outperform the best few-layer BP devices in both mobility and on/off ratio, and show comparable mobility but much higher on/off ratio than thin BP devices. Data points indexed are taken from the following: data 1 from ref. 41, data 2 from ref. 28, data 3 from ref. 42, data 4 from ref. 43, data 5 from ref. 27, data 6 from ref. 42, data 7 from ref. 20, data 8 from ref. 44, data 9 from ref. 42, data 10 from ref. 45, data 11 from ref. 46, data 12 from ref. 47, data 13 from ref. 48, data 14 from ref. 49, data 15 from ref. 50, data 16 from ref. 46, data 17 from ref. 19, data 18 from ref. 46, data 19 from ref. 51, data 20 from ref. 52, data 21 from ref. 53, data 22 from ref. 54, data 23 from ref. 55, data 24 from ref. 56, data 25 from ref. 15, data 26 from ref. 26.



#### Extended Data Figure 6 | Lateral BP-MPMS heterojunction.

**a**, Photoluminescence mapping (at 553 nm) of a lateral BP-MPMS heterostructure to highlight the MPMS part. Scale bar: 3 μm. The signal in the electrode area is due to a scattering-induced background. **b**, The corresponding Raman spectral mapping centred at 438 cm<sup>-1</sup> to show the main BP region with stronger Raman signal. Scale bar: 3 μm. **c**, SEM

image to show the lateral BP-MPMS heterojunction device. Scale bar: 3 μm. **d**, Schematic illustration of a lateral BP-MPMS heterojunction. **e**, Band diagram of the BP-MPMS heterojunction. **f**, The typical diode characteristics of a lateral BP-MPMS heterojunction; inset: optical microscope image of the corresponding BP-MPMS heterojunction. Scale bar: 3 μm.



**Extended Data Figure 7 | XRD patterns of MACMS obtained from six additional 2DACs.** **a**, XRD pattern of WSe<sub>2</sub> and WSe<sub>2</sub>/CTAB superlattice verifying the interlayer distance expansion from 6.43 Å (13.76°) of WSe<sub>2</sub> (002) peak (black) to 15.20 Å (5.81°) of WSe<sub>2</sub>/CTAB superlattice (002) peak (red). **b**, XRD pattern of SnSe and SnSe/CTAB superlattice demonstrating the interlayer distance expansion from 5.74 Å (31.16°) of SnSe (004) peak (black) to 15.62 Å (5.65°) of SnSe/CTAB superlattice (002) peak (red). **c**, XRD pattern of GeS and GeS/CTAB superlattice showing the interlayer distance expansion from 5.18 Å (34.58°) of GeS (004) peak (black) to 15.76 Å (5.60°) of GeS/CTAB superlattice (002) peak

(red). **d**, XRD pattern of NbSe<sub>2</sub> and NbSe<sub>2</sub>/CTAB superlattice revealing the interlayer distance expansion from 6.18 Å (14.31°) of NbSe<sub>2</sub> (002) peak (black) to 14.85 Å (5.95°) of NbSe<sub>2</sub>/CTAB superlattice (002) peak (red). **e**, XRD pattern of Bi<sub>2</sub>Se<sub>3</sub> and Bi<sub>2</sub>Se<sub>3</sub>/CTAB superlattice exhibiting the interlayer distance expansion from 14.16 Å (18.78°) of Bi<sub>2</sub>Se<sub>3</sub> (006) peak (black) to 23.07 Å (11.49°) of Bi<sub>2</sub>Se<sub>3</sub>/CTAB superlattice (006) peak (red). **f**, XRD pattern of In<sub>2</sub>Se<sub>3</sub> and In<sub>2</sub>Se<sub>3</sub>/CTAB superlattice indicating the interlayer distance expansion from 9.50 Å (18.67°) of the In<sub>2</sub>Se<sub>3</sub> (004) peak (black) to 15.40 Å (11.48°) of the In<sub>2</sub>Se<sub>3</sub>/CTAB superlattice (004) peak (red).



Extended Data Table 1 | Key characteristics of MPMS and recently reported few-layer BP

<b>Ref. #: BP thickness, passivation approach</b>	<b>bandgap (eV)</b>	<b>mobility (cm<sup>2</sup>/V/s)</b>	<b>on/off ratio</b>	<b>electrical stability (hour)</b>
<b>This work (MPMS)</b>	<b>2.26</b>	<b>328</b>	<b>1.9E7</b>	<b>300+</b>
<i>Ref. 19 : 10 nm, aryl diazonium passivation</i>	n.a.	~ 100	<b>1E6</b>	83+
<i>Ref. 27: 5 nm, BN passivated in inert air</i>	n.a.	118 (200K)	1E5	48+
<i>Ref. 26: 8 nm, BN encapsulated</i>	n.a.	n.a.	2E5	150+
<i>Ref. 28: 4.5 nm, BN passivated;</i>	n.a.	63	100	n.a.
<i>Ref. 21: monolayer</i>	1.73	n.a.	n.a.	n.a.
<i>Ref. 20: 5 nm, not passivated</i>	1.45	<b>286</b>	1E5	n.a.
<i>Ref. 18: Monolayer, ALD Al<sub>2</sub>O<sub>3</sub> passivated;</i>	<b>1.84</b>	n.a.	n.a.	144 + (PL)
<i>Ref. 47: 8.9 nm, ALD AlO<sub>x</sub> passivated</i>	n. a.	74	1e3	<b>175</b>

Compared with representative few-layer (less than 5 nm) phosphorene reported in the past 3 years, the MPMS allows access to intrinsic monolayer phosphorene characteristics including higher optical bandgap, higher few-layer mobility, higher on/off ratio and extraordinary stability.

# CaSiO<sub>3</sub> perovskite in diamond indicates the recycling of oceanic crust into the lower mantle

F. Nestola<sup>1</sup>, N. Korolev<sup>2,3</sup>, M. Kopylova<sup>2</sup>, N. Rotiroti<sup>4</sup>, D. G. Pearson<sup>5</sup>, M. G. Pamato<sup>6</sup>, M. Alvaro<sup>7</sup>, L. Peruzzo<sup>8</sup>, J. J. Gurney<sup>9</sup>, A. E. Moore<sup>10</sup> & J. Davidson<sup>11</sup>

Laboratory experiments and seismology data have created a clear theoretical picture of the most abundant minerals that comprise the deeper parts of the Earth's mantle. Discoveries of some of these minerals in 'super-deep' diamonds—formed between two hundred and about one thousand kilometres into the lower mantle—have confirmed part of this picture<sup>1–5</sup>. A notable exception is the high-pressure perovskite-structured polymorph of calcium silicate (CaSiO<sub>3</sub>). This mineral—expected to be the fourth most abundant in the Earth—has not previously been found in nature. Being the dominant host for calcium and, owing to its accommodating crystal structure, the major sink for heat-producing elements (potassium, uranium and thorium) in the transition zone and lower mantle, it is critical to establish its presence. Here we report the discovery of the perovskite-structured polymorph of CaSiO<sub>3</sub> in a diamond from South African Cullinan kimberlite. The mineral is intergrown with about six per cent calcium titanate (CaTiO<sub>3</sub>). The titanium-rich composition of this inclusion indicates a bulk composition consistent with derivation from basaltic oceanic crust subducted to pressures equivalent to those present at the depths of the uppermost lower mantle. The relatively 'heavy' carbon isotopic composition of the surrounding diamond, together with the pristine high-pressure CaSiO<sub>3</sub> structure, provides evidence for the recycling of oceanic crust and surficial carbon to lower-mantle depths.

A key goal of solid-Earth geosciences is to establish the mineralogy of the Earth's mantle throughout its depth, which acts as a primary control on mantle dynamics and chemistry. Diamonds are unique in this regard because they provide access to the deepest intact material from the Earth's interior through the minerals contained within their volumes. Over the past three decades, a growing number of studies have used a class of diamonds known as super-deep diamonds to study mantle processes in the deep sublithospheric mantle, the transition zone and the lower mantle<sup>1–6</sup>. Early studies<sup>1–3</sup> suggest that some of the assemblages included within super-deep diamonds represent samples of the lower mantle and the transition zone that variably retrogressed to lower pressures. Later studies indicate that some of these assemblages and minerals might originate from shallower depths<sup>7,8</sup>, although still beneath the lithosphere.

The most common minerals found within super-deep diamonds are ferropericlase [(Mg,Fe)O] and CaSiO<sub>3</sub> (refs 1–3, 9). Ferropericlase is stable at most pressure and temperature conditions in the mantle; therefore, when found as a single inclusion within diamond, this mineral cannot be considered an unambiguous indicator of a super-deep origin<sup>7</sup>.

The CaSiO<sub>3</sub> phase found within super-deep diamonds typically has the crystal structure of walsstromite (BaCa<sub>2</sub>Si<sub>3</sub>O<sub>9</sub>)<sup>1,6,8,9</sup>. Perovskite-structured CaSiO<sub>3</sub> (Ca-Pv) is considered one of the most important components in the Earth's lower mantle, comprising approximately

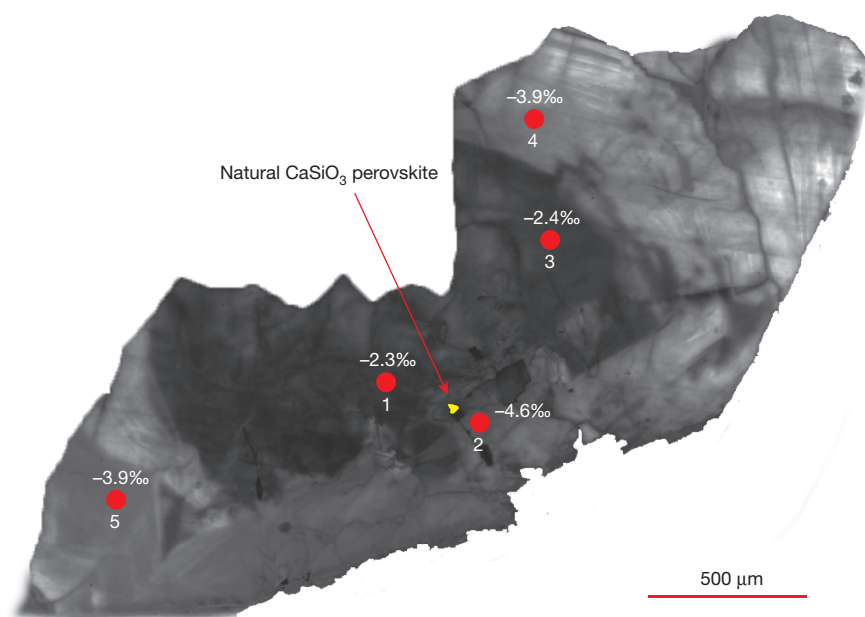
7% of the peridotitic mantle and about 23% of the volume of a subducted mid-ocean ridge basalt slab<sup>9–11</sup>. As such, it is likely to be the fourth most abundant terrestrial mineral. Within the peridotitic lower mantle, Ca-Pv is the dominant sink for Ca and for incompatible elements, including the key heat-producing elements K, U and Th (ref. 12). However, Ca-Pv has so far never been found in nature and even high-pressure laboratory experiments have failed to quench it to a metastable phase at the conditions of the Earth's surface. Although early studies<sup>1–4</sup> of super-deep diamonds make a clear case for the presence of Ca-Pv in the transition zone and lower mantle, the structure of this phase was either undetermined or documented to be the lower-pressure polymorph—CaSiO<sub>3</sub> walsstromite—and interpreted as a back-transformation of perovskite-structured CaSiO<sub>3</sub>. The phase transformation from Ca-Pv to CaSiO<sub>3</sub> walsstromite would require a volume change<sup>13</sup> of about 28%, which is impossible for diamond to accommodate owing to its extremely high bulk modulus<sup>14</sup>. The absence of healed fractures in the diamond host reported in ref. 13 implies that CaSiO<sub>3</sub> walsstromite is unlikely to represent inverted Ca-Pv. Plastic deformation of the diamond lattice could accommodate some of the volume change necessary for the phase transformation of the inclusion. However, although plastic deformation in super-deep diamonds has been well documented<sup>15</sup> and is expected to be substantial, it has never been quantified. Some super-deep diamonds with documented phase assemblages that include ferropericlase, enstatite (inverted bridgmanite) or CaSiO<sub>3</sub> walsstromite probably originate from lower-mantle depths<sup>1–4,9</sup>, but ambiguity remains. Therefore, finding an un-retrogressed Ca-Pv would provide confirmation of lower-mantle sampling by super-deep diamonds.

In this study we investigated an inclusion within a diamond from the Cullinan mine in the Gauteng province of South Africa. The Cullinan kimberlite is a group I kimberlite; that is, its chemistry and Sr, Nd and Hf isotope signatures are thought to reflect a melt source beneath the lithospheric mantle, within the Earth's convecting mantle<sup>16</sup>. The Cullinan mine is renowned for producing exceptionally large diamonds (such as the 3,107-carat Cullinan diamond<sup>6,17</sup>), most of which have been suggested to be super-deep diamonds<sup>6</sup>.

The diamond examined here has a 31 μm × 26 μm × 10 μm CaSiO<sub>3</sub> inclusion, which was exposed by polishing. X-ray diffraction, Raman spectroscopy and electron backscatter diffraction (EBSD) reveal the CaSiO<sub>3</sub> in this inclusion to have a perovskite structure. To our knowledge, this represents the only finding of non-reverted Ca-Pv in nature and the first Ca-Pv, including those synthesized in the laboratory, to preserve its high-pressure structure at the surface of the Earth.

Cathodoluminescence imaging of the host diamond surrounding the Ca-Pv inclusion (Fig. 1) reveals multiple growth zones and a complex internal structure, typical of super-deep diamonds<sup>4,18</sup>. Fourier

<sup>1</sup>Dipartimento di Geoscienze, Università degli Studi di Padova, Via Giovanni Gradeno 6, I-35131 Padova, Italy. <sup>2</sup>Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. <sup>3</sup>Institute of Precambrian Geology and Geochronology RAS, 199034 St Petersburg, Russia. <sup>4</sup>Dipartimento di Scienze della Terra, Università degli Studi di Milano, Via Botticelli 23, I-20133 Milano, Italy. <sup>5</sup>Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada. <sup>6</sup>Department of Earth Sciences, University College London, Gower Street, London WC1E 6BT, UK. <sup>7</sup>Department of Earth and Environmental Sciences, University of Pavia, Via Ferrata 1, I-27100 Pavia, Italy. <sup>8</sup>CNR-Istituto di Geoscienze e Georisorse, Sezione di Padova, Via Giovanni Gradeno 6, I-35131 Padova, Italy. <sup>9</sup>University of Cape Town, Cape Town, South Africa. <sup>10</sup>Rhodes University, Grahamstown, South Africa. <sup>11</sup>Petra Diamonds, Bryanston, South Africa.



**Figure 1 | Cathodoluminescence image and carbon isotopic composition of the diamond containing the Ca-Pv inclusion.** The Ca-Pv inclusion is shown in yellow. Carbon isotopic compositions measured at five locations (red circles) are expressed as  $\delta^{13}\text{C}$  values.

transform infrared (FTIR) spectroscopy (Extended Data Fig. 1) of the diamond host indicates a nitrogen content of 34 p.p.m., with 97% in the B-aggregated form; that is, the diamond host is type IaB. The low nitrogen content and very high level of B aggregation are typical characteristics of super-deep diamonds<sup>4,19</sup>, indicating prolonged exposure to the high temperatures that are prevalent at transition-zone and lower-mantle depths.

The chemical composition of the Ca-Pv inclusion, determined by electron microprobe analysis, is almost pure  $\text{CaSiO}_3$  ( $\text{Ca}_{0.98}\text{Si}_{0.98}\text{O}_3$ ), with minor impurities of Ti, Al, Fe and Mg totalling 0.04 atoms per formula unit (Extended Data Table 1).

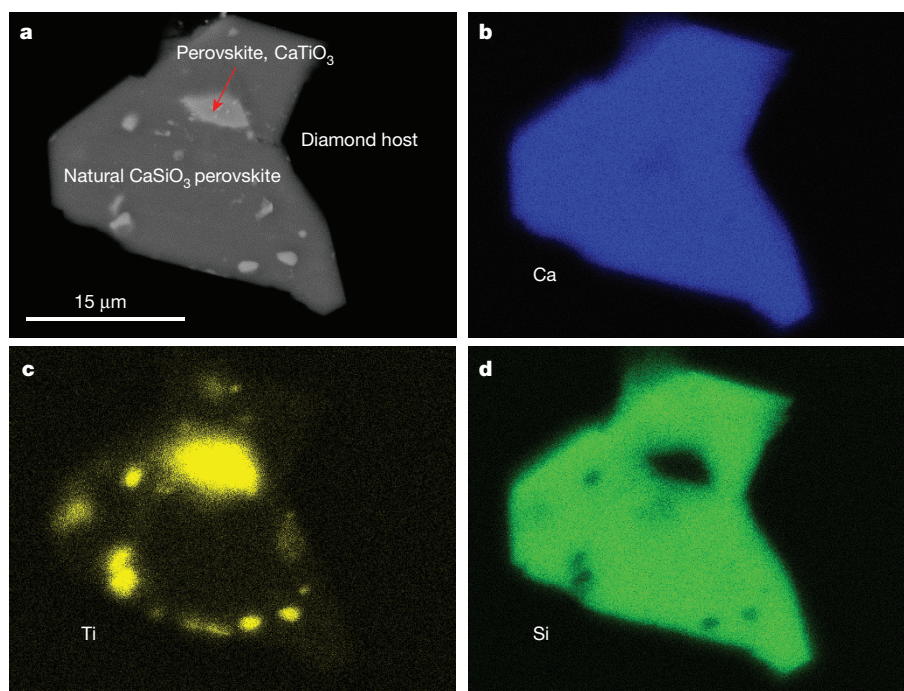
Backscattered-electron imaging and energy-dispersive X-ray spectroscopy (EDS) element maps (Fig. 2) show that the Ca-Pv crystal includes 14 irregular areas of  $\text{CaTiO}_3$  perovskite with sizes between 1  $\mu\text{m}$  and 7–8  $\mu\text{m}$  and an approximate stoichiometry of  $\text{Ca}(\text{Ti}_{0.92}\text{Si}_{0.07}\text{Al}_{0.02})\text{O}_3$ . The texture, size and abundance of these  $\text{CaTiO}_3$  intergrowths are very similar to those of inclusions reported in  $\text{CaSiO}_3$  walsstromite phases in super-deep diamonds<sup>20</sup> from Juina, Brazil. The exposed surface of our Ca-Pv inclusion makes accurate estimation of its bulk composition difficult, but image analysis indicates that the host crystal in bulk may contain up to 6% by volume  $\text{CaTiO}_3$ .  $\text{CaTiO}_3$  perovskite is a common mineral in nature and remains stable well into the lower mantle<sup>21</sup>. By contrast, a Ca-Pv sample that retains its perovskite structure at room temperature and pressure has no experimentally synthesized analogues, unless a considerable amount of  $\text{CaTiO}_3$  (about 34 mol%; ref. 21) is dissolved within its structure, far more than the  $\text{CaTiO}_3$  component observed here. However, our discovery of natural Ca-Pv with less than 2 mol%  $\text{CaTiO}_3$  in the  $\text{CaSiO}_3$ -rich portion of the inclusion indicates that, unlike experiments, nature must provide pressure–temperature–time pathways that are capable of preserving this metastable phase.

As stated above, X-ray diffraction data show that the  $\text{CaSiO}_3$  inclusion has a perovskite structure. The small size of the inclusion (thickness  $\leq 10 \mu\text{m}$ , as estimated by confocal Raman spectroscopy) and its entrapment within the diamond host resulted in only a limited number of measured diffraction reflections ( $n = 91$ ), of which only nine were unique (Extended Data Table 2). All of the 91 reflections were used to refine the Ca-Pv unit-cell parameters:  $a = 5.397 \pm 0.004 \text{ \AA}$ ,  $b = 5.404 \pm 0.004 \text{ \AA}$ ,  $c = 7.646 \pm 0.004 \text{ \AA}$ , volume  $223.0 \pm 0.03 \text{ \AA}^3$

However, alternative unit-cell refinements using other numerical approaches could provide considerably different (by more than 1%) unit-cell parameters owing to the relatively poor accuracy and precision with which the spacings between crystal planes ( $d$  spacings) were measured in this study. These relatively large uncertainties are typical when studying minerals of this size and arise not only from the limited number of reflections, but also because the measurements were performed using an area detector, which provides lower precision in  $d$ -spacing determination than a point detector. Such relatively large uncertainty on the cell parameters makes any comparison with the unit-cell volume of  $\text{CaTiO}_3$  perovskite unreliable, although we can establish that the structures of  $\text{CaTiO}_3$  perovskite and Ca-Pv are very similar. Ewald projections along the three crystallographic axes (Fig. 3a) indicate an orthorhombic unit cell. The unit cell and the chemical composition confirm that the mineral is Ca-Pv. Recent numerical simulations on ‘host–inclusion’ systems<sup>22</sup> indicate that an inclusion partly exposed to atmospheric pressure loses only a portion of its residual pressure, depending on the elastic properties of both the host mineral and the inclusion. The Ca-Pv inclusion studied here is partly exposed at the diamond surface, but with two-thirds of its volume still buried in the diamond host. Thus, any measurements on this grain would be affected by some residual pressure still acting on the inclusion, which in turn affects the X-ray diffraction data and Raman spectra.

Raman spectra (Fig. 3b) of the inclusion show that the spectrum of the  $\text{CaTiO}_3$  perovskite is in excellent agreement with Raman data for  $\text{CaTiO}_3$  perovskite from the RRUFF database<sup>23</sup> (Extended Data Fig. 2). The  $\text{CaSiO}_3$  and  $\text{CaTiO}_3$  spectra are similar. Small differences are evident because of the presence of two Raman peaks for the  $\text{CaSiO}_3$  spectrum, which could belong to the lower-pressure  $\text{CaSiO}_3$  polymorph<sup>23</sup> wollastonite-2M. This wollastonite polymorph is not stable at pressures higher than 3 GPa along a mantle geotherm<sup>24</sup>, well below the diamond stability field. Therefore, its presence is probably due to minor partial inversion of the Ca-Pv phase caused by the polishing of the sample to expose the inclusion, as reported previously<sup>25</sup>.

EBSD measurements conducted on several areas of the grain provide no evidence of amorphous portions (Fig. 4). The EBSD pattern of the  $\text{CaSiO}_3$  area (red circle), shown as the non-indexed EBSD pattern in Fig. 4b, is complex and could not be indexed by a single phase. The observed pattern can be indexed by using a combination of



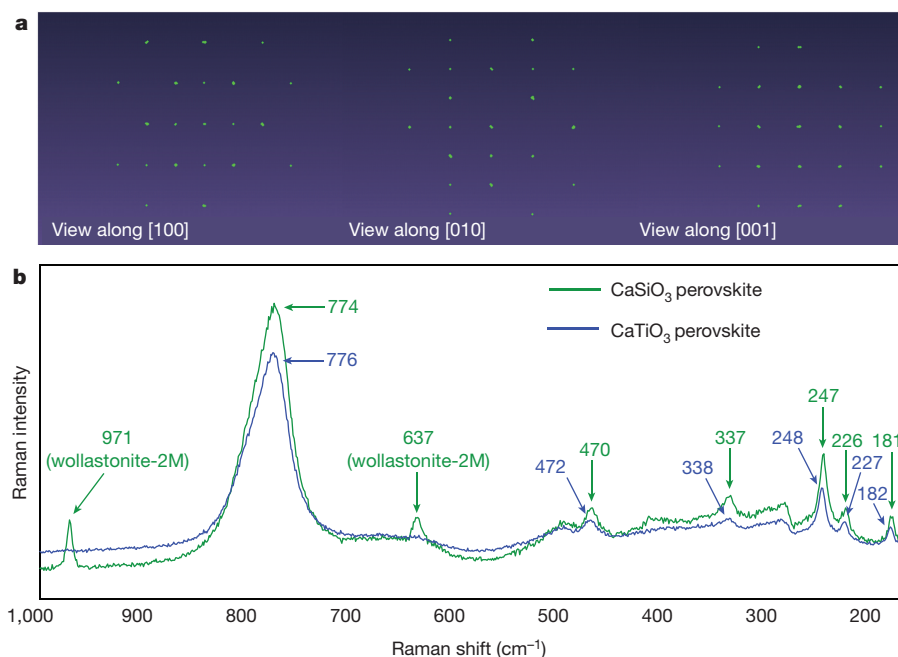
**Figure 2 | Backscattered-electron image of the Ca-Pv inclusion and energy-dispersive X-ray spectroscopy elemental maps. a**, Backscattered-electron image of the Ca-Pv inclusion (dark grey) surrounded by the diamond host (black), showing smaller inclusions of  $\text{CaTiO}_3$  perovskite

(light grey). **b–d**, Energy-dispersive X-ray spectroscopy elemental maps of Ca (**b**), Ti (**c**) and Si (**d**). The colour intensity (black within the grain outline through to saturation in a specific colour) is proportional to the element concentration.

reference EBSD patterns for  $\text{CaTiO}_3$  perovskite (Fig. 4c) and  $\text{CaSiO}_3$  wollastonite-2M (Fig. 4d), again confirming that  $\text{CaSiO}_3$  is present in this diamond with a perovskite-type structure. Because EBSD measures surface responses (within tens of nanometres from the surface), this signal can only come from the  $\text{CaSiO}_3$  phase.

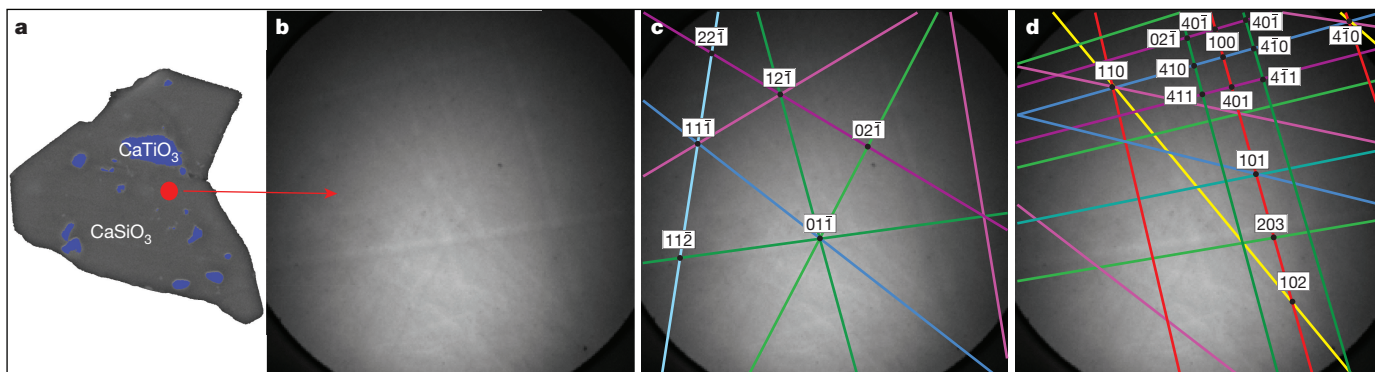
We suggest that the natural Ca-Pv found trapped within the super-deep diamond was a result of the unmixing of the high-pressure solid solution  $\text{Ca}(\text{Ti},\text{Si})\text{O}_3$ . If the two phases exsolved from a homogeneous bulk composition, this phase would contain about 3.9%  $\text{TiO}_2$ .

We estimate that the stoichiometry of the original phase composition was  $(\text{Ca}_{0.98}\text{Mg}_{0.01}\text{Fe}_{0.01})(\text{Si}_{0.93}\text{Ti}_{0.06}\text{Al}_{0.01})\text{O}_3$ . This composition is consistent with that of  $\text{CaSiO}_3$  samples crystallized in experiments<sup>26</sup> from a mid-ocean ridge basalt (MORB)-like bulk composition at about 26 GPa and is similar to that of  $\text{CaSiO}_3$  walsstromite and  $\text{CaTiO}_3$  intergrowths found within Juina super-deep diamonds; these intergrowths have been suggested to originate from basalt-like compositions subducted to lower-mantle depths that later retrogressed during their ascent to the Earth's surface<sup>20</sup>. The preservation of the high-pressure perovskite



**Figure 3 | Ewald projections of X-ray diffraction data and Raman spectroscopy results. a**, Ewald projections along three different orientations for the Ca-Pv. **b**, Baseline-corrected Raman spectrum of the Ca-Pv inclusion compared with that of the  $\text{CaTiO}_3$ -perovskite intergrowth (Fig. 2a).





**Figure 4 | EBSD images of the Ca-Pv inclusion in diamond.** **a**, Location (indicated by the red filled circle) from which the EBSD images were obtained. **b**, Non-indexed EBSD pattern relative to the  $\text{CaSiO}_3$  inclusion. **c**, **d**, EBSD patterns indexed with the reference patterns of  $\text{CaTiO}_3$  (**c**) and  $\text{CaSiO}_3$  wollastonite-2M (**d**). The coloured lines in **c** represent the EBSD

indexed pattern of  $\text{CaTiO}_3$ , whereas the coloured lines in **d** represent the EBSD indexed pattern of wollastonite-2M. The numbers reported in both the figures at the intersections between the coloured lines represent the zone axes.

structure in the case of the Cullinan inclusion supports the derivation of such compositions from lower-mantle depths.

The possible subducted basaltic protolith origin of the Cullinan Ca-Pv inclusion suggests that we might expect to observe some evidence of a crustal parentage in the carbon isotopic composition of the host diamond (Fig. 1; Extended Data Table 3), which has  $\delta^{13}\text{C}$  values ranging from  $-2.3\text{‰}$  to  $-4.6\text{‰}$ , where  $\delta^{13}\text{C} = (^{13}\text{C}/^{12}\text{C})_{\text{sample}} / (^{13}\text{C}/^{12}\text{C})_{\text{PDB}} - 1$  (PDB, Pee Dee Belemnite reference material). The core region of the diamond, defined by cathodoluminescence imaging (Fig. 1), contains the Ca-Pv inclusion and has an average  $\delta^{13}\text{C}$  value of  $-2.3\text{‰} \pm 0.5\text{‰}$ , considerably lower than the typical upper-mantle value<sup>27</sup> of  $-5.5\text{‰}$ . By contrast, the outer-rim region of the diamond has a composition (mean  $\delta^{13}\text{C}$  of  $-4.1\text{‰} \pm 0.5\text{‰}$ ) that is closer to  $-5.5\text{‰}$ . Crustal carbon reservoirs have carbon isotopic compositions that are both ‘heavier’ and ‘lighter’ than the typical upper-mantle value. While ‘isotopically light’ carbon compositions ( $\delta^{13}\text{C} < -25\text{‰}$ ) have been found in super-deep diamonds from Juina, which are thought to be derived from subducted basalt protoliths<sup>20,28</sup>, ‘isotopically heavy’ ( $-3\text{‰}$  to  $-0.5\text{‰}$ ) carbon compositions, such as those measured in the core of the studied diamond, have also been reported in super-deep diamonds from Brazil (Sao Luis and Juina) and Guinea (Kankan)<sup>18–20</sup>. If the  $\delta^{13}\text{C}$  value of  $-2.3\text{‰}$  is compared with the median value ( $-4.91\text{‰}$ ) of 1,473 published analyses of lithospheric diamonds containing peridotitic inclusions—a group of diamonds usually accepted to be minimally affected by subduction<sup>27</sup>—it is found to be an outlier, beyond three times the median absolute deviation. Such anomalously high carbon isotopic compositions are thought to reflect a greater influence of subducted carbonate in the fluid that formed these super-deep diamonds<sup>18,19</sup>. The carbon isotope compositions of the rim of the Cullinan diamond (Fig. 1) may represent an overgrowth that developed under upper-mantle conditions or from a distinct source of carbon in the lower mantle. Regardless, the high  $\delta^{13}\text{C}$  values of the portion of the diamond that contains the Ca-Pv inclusion supports the premise that it originates from a subducted basaltic protolith.

Our discovery of Ca-Pv in a super-deep diamond firmly establishes this phase as a component of the Earth’s deep mantle, confirming previous suggestions<sup>1–4,9</sup> that lower-pressure  $\text{CaSiO}_3$  polymorphs included in these diamonds may represent retrogressed Ca-Pv. The estimated original bulk composition of the Cullinan Ca-Pv inclusion is consistent with compositions that are stable in subducted oceanic basalt protoliths at about 26 GPa, in the uppermost lower mantle<sup>26</sup>. Our finding thus confirms the expectation from calculations<sup>10</sup> and high-pressure experiments<sup>21,25</sup>, that Ca-Pv is the main Ca-bearing phase in the lower mantle in both basic and ultrabasic compositions, reaching up to 23 vol% in MORB-like compositions<sup>26</sup>. The combined bulk composition of the Ca-Pv phase found here provides overwhelming evidence

of the return of recycled oceanic crust into the Earth’s lower mantle<sup>20</sup>, whereas the relatively high carbon isotopic composition of the diamond in contact with the inclusion indicates the subduction of crustal carbon to lower mantle depths.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 27 July 2017; accepted 19 January 2018.**

- Joswig, W., Stachel, T., Harris, J. W., Baur, W. & Brey, G. P. New Ca-silicate inclusions in diamonds – tracers from the lower mantle. *Earth Planet. Sci. Lett.* **173**, 1–6 (1999).
- Harte, B. in *Mantle Petrology: Field Observations and High-Pressure Experimentation* Vol. 6 (eds Fei, Y. et al.) 125–153 (Geochemical Society, 1999).
- Stachel, T., Harris, J. W., Brey, G. P. & Joswig, W. Kankan diamonds (Guinea) II: lower mantle inclusion paragenesis. *Contrib. Mineral. Petrol.* **140**, 16–27 (2000).
- Hayman, P. C., Kopylova, M. G. & Kaminsky, F. V. Lower mantle diamonds from Rio Soriso (Juina area, Mato Grosso, Brazil). *Contrib. Mineral. Petrol.* **149**, 430–445 (2005).
- Pearson, D. G. et al. Hydrous mantle transition zone indicated by ringwoodite included within diamond. *Nature* **507**, 221–224 (2014).
- Smith, E. M. et al. Large gem diamonds from metallic liquid in Earth’s deep mantle. *Science* **354**, 1403–1405 (2016).
- Brey, G. P., Bulatov, V., Girs, A., Harris, J. W. & Stachel, T. Ferropericlasite—a lower mantle phase in the upper mantle. *Lithos* **77**, 655–663 (2004).
- Thomson, A., Walter, M. J., Kohn, S. C. & Brooker, R. A. Slab melting as a barrier to deep carbon subduction. *Nature* **529**, 76–79 (2016).
- Harte, B. & Hudson, N. F. C. Mineral associations in diamonds from the lowermost upper 254 mantle and uppermost lower mantle. In *Proc. of the 10th International Kimberlite Conference* Vol. 1 (eds Pearson, D. G. et al.) 235–253 (Springer, 2013).
- Stixrude, L. & Lithgow-Bertelloni, C. Geophysics of chemical heterogeneity in the mantle. *Annu. Rev. Earth Planet. Sci.* **40**, 569–595 (2012).
- Ringwood, A. E. *Composition and Petrology of the Earth’s Mantle* (McGraw-Hill, 1975).
- Corgne, A. & Wood, B. J. Trace element partitioning and substitution mechanisms in calcium perovskites. *Contrib. Mineral. Petrol.* **149**, 85–97 (2005).
- Anzolini, C. et al. Depth of formation of  $\text{CaSiO}_3$ -walsstromite included in super-deep diamonds. *Lithos* **265**, 138–147 (2016).
- Angel, R. J., Alvaro, M., Nestola, F. & Mazzucchelli, M. L. Diamond thermoelastic properties and implications for determining the pressure of formation of diamond-inclusion systems. *Russ. Geol. Geophys.* **56**, 211–220 (2015).
- Cayzer, N. J., Odake, S., Harte, B. & Kagi, H. Plastic deformation of lower mantle diamonds by inclusion phase transformations. *Eur. J. Mineral.* **20**, 333–339 (2008).
- Nowell, G. M. et al. Hf isotope systematics of kimberlites and their megacrysts: new constraints on their source region. *J. Petrol.* **45**, 1583–1612 (2004).
- Moore, A. D. The origin of large irregular gem-quality type II diamonds and the rarity of blue type IIb varieties. *S. Afr. J. Geol.* **117**, 219–236 (2014).
- Palot, M., Pearson, D. G., Stern, R. A., Stachel, T. & Harris, J. W. Isotopic constraints on the nature and circulation of deep mantle C-H-O-N fluids: carbon and nitrogen systematics within ultra-deep diamonds from Kankan (Guinea). *Geochim. Cosmochim. Acta* **139**, 26–46 (2014).

19. Stachel, T., Harris, J. W., Aulbach, S. & Deines, P. Kankan diamonds (Guinea) III:  $\delta^{13}\text{C}$  and nitrogen characteristics of deep diamonds. *Contrib. Mineral. Petrol.* **142**, 465–475 (2002).
20. Walter, M. J. *et al.* Deep mantle cycling of oceanic crust: evidence from diamonds and their mineral inclusions. *Science* **334**, 54–57 (2011).
21. Kubo, A., Suzuki, T. & Akaogi, M. High pressure phase equilibria in the system  $\text{CaTiO}_3\text{--CaSiO}_3$ : stability of perovskite solid solutions. *Phys. Chem. Miner.* **24**, 488–494 (1997).
22. Mazzucchelli, M. L. *et al.* Elastic geothermobarometry: corrections for the geometry of the host-inclusion system. *Geology* <https://doi.org/10.1130/G39807.1> (2018).
23. Lafuente, B., Downs, R.T., Yang, H. & Stone, N. in *Highlights in Mineralogical Crystallography* (eds Armbruster, T. & Danisi, R. M.) 1–30 (De Gruyter, 2016).
24. Gasparik, T., Wolf, L. & Smith, C. M. Experimental determination of phase relations in the  $\text{CaSiO}_3$  system from 8 to 15 GPa. *Am. Mineral.* **79**, 1219–1222 (1994).
25. Ringwood, A. E. & Major, A. Synthesis of majorite and other high pressure garnets and perovskites. *Earth Planet. Sci. Lett.* **12**, 411–418 (1971).
26. Hirose, K. & Fei, Y. Subsolidus and melting phase relations of basaltic composition in the uppermost lower mantle. *Geochim. Cosmochim. Acta* **66**, 2099–2108 (2002).
27. Cartigny, P., Palot, M., Thomassot, E. & Harris, J. W. Diamond formation: a stable isotope perspective. *Annu. Rev. Earth Planet. Sci.* **42**, 699–732 (2014).
28. Burnham, A. D. *et al.* Stable isotope evidence for crustal recycling as recorded by superdeep diamonds. *Earth Planet. Sci. Lett.* **432**, 374–380 (2015).

**Acknowledgements** We thank M. Regier for proofreading the paper. F.N. is supported by the European Research Council (ERC) Starting Grant number 307322. M.K.'s work and sample collection was possible thanks to an NSERC Discovery grant. N.K. acknowledges funding from the Dr. Eduard Gübelin

Association through a 2015 research scholarship. D.G.P. was funded by an NSERC CERC award. M.A. was supported by the ERC under the European Union's Horizon 2020 research and innovation programme (grant 714936) 'TRUE DEPTHS' and by the SIR-MIUR grant (RBSI140351) 'MILE DEEP'. We thank L. Litti and M. Meneghetti of the Laboratory of Nanostructures and Optics of the Department of Chemical Sciences, University of Padova for their help in acquiring and interpreting the Raman data. F.N. and D.G.P. were supported by the Deep Carbon Observatory. M.G.P. was supported by NERC grant NE/M015181/1.

**Author Contributions** F.N. conceived the study, wrote the initial manuscript and performed X-ray diffraction and micro-Raman measurements. N.K. found the mineral, made original mineral identifications on a confocal Raman spectrometer, performed microprobe and cathodoluminescence measurements, prepared samples for secondary ion mass spectrometry measurements and assisted with the manuscript preparation. M.K. supervised the study of the Cullinan diamond collection, which was acquired by J.J.G., A.E.M. and J.D., and assisted with the manuscript preparation. D.G.P. made the geochemical interpretations and led the manuscript revisions. M.G.P. assisted with the manuscript preparation and crystallographic interpretations. N.R., M.G.P. and M.A. assisted with the X-ray data interpretation. L.P. collected and interpreted the EBSD data. J.J.G., A.E.M. and J.D. designed the sampling programme.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to F.N. ([fabrizio.nestola@unipd.it](mailto:fabrizio.nestola@unipd.it)).

**Reviewer Information** *Nature* thanks B. Harte and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**Micro-Raman spectroscopy.** The Ca-Pv sample was analysed using an InVia Renishaw micro-Raman spectrometer installed at the Department of Chemical Sciences of the University of Padova. The spectra were baseline-corrected. A 632.8-nm-wavelength excitation laser was used at a power of 7 mW. The Raman spectrum of the Ca-Pv crystal was collected for 40 s using a 50 $\times$  objective with a spatial resolution of 1.1  $\mu\text{m}$  and a spectral resolution estimated to be about 3  $\text{cm}^{-1}$ . The most intense Raman peaks observed for the Ca-Pv inclusion are (in order of decreasing intensity): 774, 247, 470, 337, 181 and 226  $\text{cm}^{-1}$ .

Direct comparison between the Raman spectrum of natural Ca-Pv with that of  $\text{CaTiO}_3$ —both from the  $\text{CaTiO}_3$  inclusions (Fig. 3) and from the RRUFF database<sup>23</sup>—indicates that the two spectra are very similar. A small but important difference is the presence of limited traces of wollastonite-2M in the Raman spectrum of the natural Ca-Pv (see peaks at 971  $\text{cm}^{-1}$  and 637  $\text{cm}^{-1}$ ), which, as expected, are not evident in the spectrum of the  $\text{CaTiO}_3$  perovskite inclusions.

On the basis of the results of ref. 29, the broad Raman bands in the 650–850  $\text{cm}^{-1}$  region correspond to second-order Raman scattering and only the sharp peaks in the 200–500  $\text{cm}^{-1}$  region are first-order Raman bands. However, we considered the entire Raman spectrum of natural Ca-Pv, regardless of first- or second-order scattering, for a direct comparison with  $\text{CaTiO}_3$  perovskite and wollastonite.

The strong similarity between the Raman spectra of the  $\text{CaTiO}_3$  and  $\text{CaSiO}_3$  perovskites in Fig. 3b could indicate that the spectra are dominated by emission from a larger, underlying, unexposed volume of crystalline  $\text{CaTiO}_3$  surrounded by a matrix of amorphous  $\text{CaSiO}_3$ . This possibility can be discounted for a number of reasons. First, the partially exposed inclusion is under some stress and this will affect the Raman band shift, depending on the elastic properties of the two perovskites. More importantly, the Raman spectrum of such a hypothetical large amorphous area of  $\text{CaSiO}_3$  would be totally distinct from that measured here (Fig. 3b), and would be characterized by the presence of three very intense Raman bands at about 370  $\text{cm}^{-1}$ , 640  $\text{cm}^{-1}$  and 970  $\text{cm}^{-1}$  (depending on the pressure and temperature conditions<sup>30,31</sup>). These Raman bands are absent in the spectrum of the  $\text{CaSiO}_3$  portion of the perovskite-structured inclusion. Also, because of the confocal nature of the Raman measurements and their small spot size, they cannot be substantially influenced by the spatially associated  $\text{CaTiO}_3$  intergrowth. Last, the EBSD measurements rule out this possibility because EBSD is a surface technique (see Methods section ‘EBSD’).

**Cathodoluminescence.** The cathodoluminescence scanning electron microscopy image shown in Fig. 1 was obtained using a Philips XL 30 scanning electron microscope with a cathodoluminescence attachment consisting of a Hamamatsu R376 photomultiplier tube (EOAS UBC). The accelerating voltage was 20 keV and the electron beam current was 100  $\mu\text{A}$ .

**Infrared spectroscopy.** Infrared spectra were collected on a Nicolet 6700 FTIR spectrometer. The absorbance spectra were measured at maximum light transmission for 40 s at a spectral resolution of 0.5  $\text{cm}^{-1}$ . Background spectra were collected for 120 s before the analysis and were subtracted from each measured absorbance spectrum. The nitrogen concentration and aggregation were determined using the procedure described in ref. 32 using a spreadsheet (‘FTIR analyser 3d’) created by J. Chapman (Rio Tinto Diamonds Ltd). Preliminary processing and baseline determination were performed using the EssentialFTIR software. The analytical and processing error was  $\pm 10\%$  (1 $\sigma$ , relative error). The FTIR spectrum of the Cullinan diamond studied here is shown in Extended Data Fig. 1.

**Electron microprobe analysis.** Quantitative chemical analyses were performed at the Department of Earth, Ocean and Atmospheric Sciences of the University of British Columbia, using a fully automated CAMECA SX-50 electron microprobe operating in the wavelength-dispersion mode with the following operating conditions: excitation voltage, 15 kV; beam current, 20 nA; peak counting time, 20 s; background counting time, 10 s; actual spot diameter, 5  $\mu\text{m}$ . Data reduction was done using the PAP  $\varphi(\rho z)$  method<sup>33</sup>. The detection limits for most oxides were lower than 0.08 wt% and those for  $\text{Cr}_2\text{O}_3$ ,  $\text{MnO}_2$  and  $\text{NiO}$  were lower than 0.12 wt%. Because of the small crystal size of the natural Ca-Pv and the presence of  $\text{CaTiO}_3$  perovskite inclusions, we were able to perform only three reliable analyses; the results are reported in Extended Data Table 1. The Na and K contents were not analysed.

**Scanning electron microscopy and EDS.** We studied our sample using scanning electron microscopy and EDS to investigate the distribution of Ca, Si and Ti over the grain. We used a CamScan MX3000 electron microscope equipped with a LaB<sub>6</sub>

source, a four-quadrant solid-state backscattered-electron detector and an EDAX EDS system for micro-analysis installed at the Department of Geosciences of the University of Padova. The measurement conditions were: accelerating voltage, 20 kV; filament emission, about 13 nA; working distance (the distance between the specimen and the lowest part of the electromagnetic lens in the column of the scanning electron microscope), 27 mm. The backscattered-electron image of the grain and its EDS maps for Ca, Si and Ti are shown in Fig. 2.

**Single-crystal micro-X-ray diffraction.** Single-crystal X-ray diffraction measurements were performed at the Department of Geosciences of the University of Padova, using a Rigaku Oxford Diffraction Supernova goniometer equipped with a Dectris Pilatus 200 K area detector and a Mova X-ray micro-source (Mo K $\alpha$  radiation) operating at 50 kV and 0.8 mA. The sample-to-detector distance was 68 mm. Data reduction was performed using the CrysAlis software (Rigaku Oxford Diffraction) to obtain the Ewald projections shown in Fig. 3. The diffraction analysis results are reported in Extended Data Table 2 in comparison with those of a reference  $\text{CaTiO}_3$  single-crystal sample<sup>34</sup> with the following unit-cell parameters:  $a = 5.388 \pm 0.001$  Å,  $b = 5.447 \pm 0.001$  Å,  $c = 7.654 \pm 0.001$  Å, volume,  $224.63 \pm 0.001$  Å<sup>3</sup>.

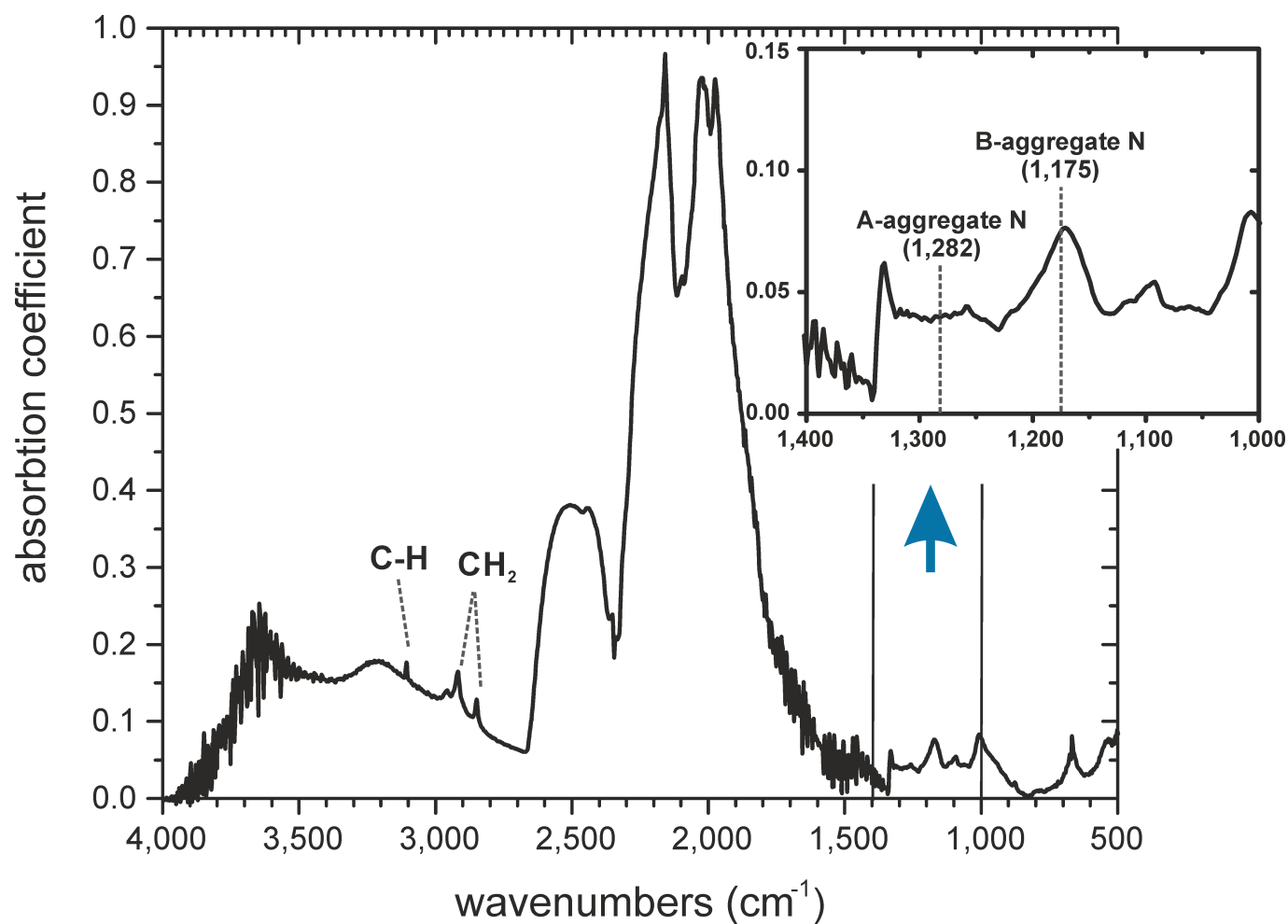
**Carbon isotope analyses.** The carbon isotope compositions ( $\delta^{13}\text{C}$ ) reported in Extended Data Table 3 were determined using a Cameca IMS 7f-GEO secondary ion mass spectrometer. The polished diamond was pressed into an indium mount with a 1-inch-diameter aluminium holder. Natural reference diamonds with  $\delta^{13}\text{C}$  values between  $-13.6\text{‰}$  ( $2\sigma = 0.3\text{‰}$ ) and  $2.6\text{‰}$  ( $2\sigma = 0.3\text{‰}$ ) were used to determine the instrumental mass fractionation and drift before and after sample analyses. The sample and reference diamonds were coated with gold (20 nm thickness). The measurements were conducted using  $^{133}\text{Cs}^+$  at 10 keV impact energy and a beam current of about 4 nA. The 15- $\mu\text{m}$ -diameter  $^{133}\text{Cs}^+$  primary-ion beam was used for pre-sputtering. During the measurements, the ion beam diameter was reduced to 5  $\mu\text{m}$ . Secondary ions of  $^{12}\text{C}$  and  $^{13}\text{C}$  were extracted at  $-9$  keV with an energy bandwidth of 90 eV. No electron-gun charge compensation was required. The  $^{13}\text{C}/^{12}\text{C}$  ratios were measured using dual Faraday cups. The mass resolving power  $M/\Delta M$  was 2,900. The  $^{12}\text{C}$  and  $^{13}\text{C}$  ions were counted for 1 s in each cycle of the 30 cycles and the total measurement time for each spot was 8 min. The standard deviation of the analysis is estimated to be about 0.4‰ to 0.5‰ at the  $2\sigma$  (95% uncertainty) level.

**EBSD.** EBSD analyses were performed at CNR-ICMATE in Padova, using a Quanta 200F FEG-ESEM system operating in high-vacuum mode with an accelerating voltage of 30 kV, emission current of 174  $\mu\text{A}$  and beam spot of 4.5  $\mu\text{m}$ , without any conductive coating. EBSD patterns were collected at a working distance of 10 mm and a specimen tilt of 75° using an EDAX Digiview EBSD system. The instrument was controlled by the OIMTM 5.31 software, which contains a large EBSD pattern database.

**Statistical analysis of carbon isotope composition.** We used a compilation of 1,473 carbon-isotope analysis datasets for diamonds containing inclusions of lithospheric peridotite paragenesis from ref. 35. We calculated the median absolute deviation<sup>36</sup> of the data using a  $b$  factor of 1.4826 and a very conservative threshold factor of 3.

**Data availability.** All relevant data are presented in Extended Data Tables 1–3 and Figs 1–4. Original spectral data and electron microprobe data are available from the corresponding author.

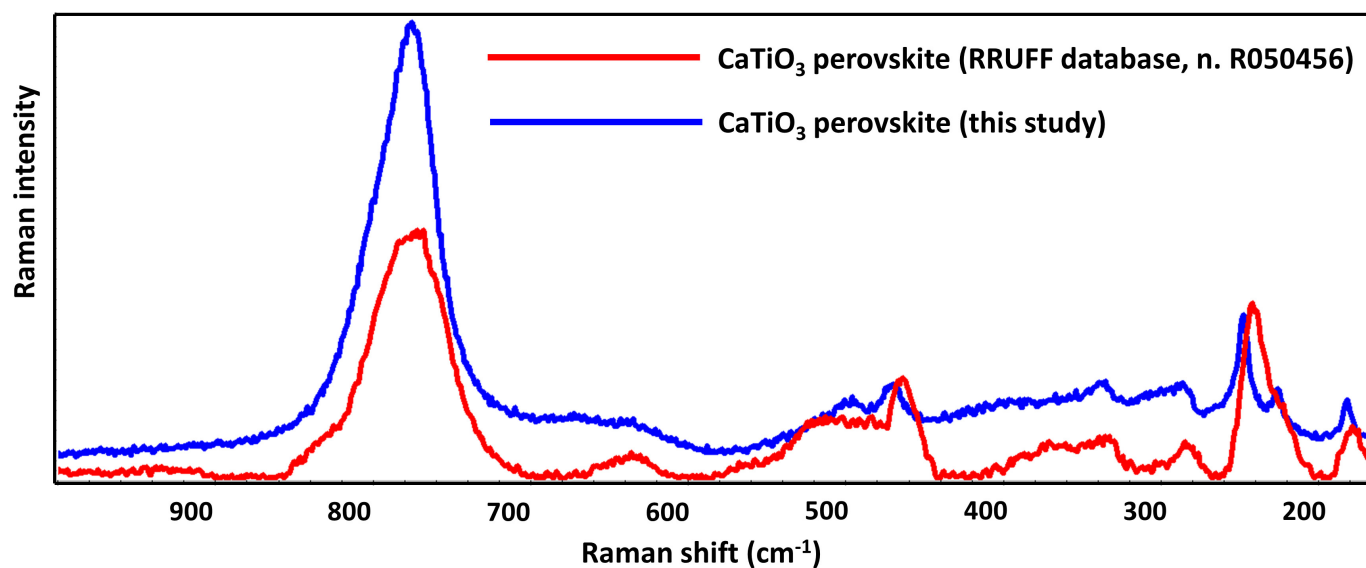
- McMillan, P. & Ross, N. The Raman spectra of several orthorhombic calcium oxide perovskites. *Phys. Chem. Miner.* **16**, 21–28 (1988).
- Yin, C. D., Okuno, M., Morikawa, H., Marumo, F. & Yamanaka, T. Structural analysis of  $\text{CaSiO}_3$  glass by X-Ray diffraction and Raman spectroscopy. *J. Non-Cryst. Solids* **80**, 167–174 (1986).
- Zerr, A., Serghiou, G. & Boehler, R. Melting of  $\text{CaSiO}_3$  perovskite to 430 kbar and first *in-situ* measurements of lower mantle eutectic temperatures. *Geophys. Res. Lett.* **24**, 909–912 (1997).
- Mendelsohn, M. J. & Milledge, H. J. Geologically significant information from routine analysis of the mid-infrared spectra of diamonds. *Int. Geol. Rev.* **37**, 95–110 (1995).
- Pouchou, J. L. & Pichoir, F. in *Microbeam Analysis* (ed. Armstrong, J. T.) 104–106 (San Francisco Press, 1985).
- Buttner, R. H. & Maslen, E. N. Electron difference density and structural parameters in  $\text{CaTiO}_3$ . *Acta Crystallogr. C* **48**, 644–649 (1992).
- Stachel, T., Harris, J. W. & Muehlenbachs, K. Sources of carbon in inclusion bearing diamonds. *Lithos* **112**, 625–637 (2009).
- Huber, P. *Robust Statistics* 107–108 (Wiley, 1981).



**Extended Data Figure 1 | Baseline-corrected FTIR absorption spectrum of the diamond containing the Ca-Pv inclusion.** The inset shows the absorption peaks of two types of diamond defect: A-aggregate N, in which a pair of nitrogen atoms substitute carbon atoms (the 'A-centre'), and

B-aggregate N, in which four nitrogen atoms replace carbon atoms around a carbon vacancy (the 'B-centre'). The values in parentheses give the theoretical peak positions (in  $\text{cm}^{-1}$ ).





Extended Data Figure 2 | Comparison between the Raman spectrum of CaTiO<sub>3</sub> measured in this work (blue) and that reported in the RRUFF database (red). The spectrum reported in the RRUFF database (card number R050456) is from ref. 23.

Extended Data Table 1 | Results of the chemical analysis of the Ca-Pv

Oxide	Wt.% (Ave)	Range	Cations (O=3)	Stand. Dev.	Probe Standard
SiO <sub>2</sub>	50.26	50.13-50.38	0.984	0.10	CaMgSi <sub>2</sub> O <sub>6</sub> , SiK $\alpha$ , TAP
TiO <sub>2</sub>	0.35	0.22-0.47	0.005	0.09	TiO <sub>2</sub> , TiK $\alpha$ , PET
Al <sub>2</sub> O <sub>3</sub>	0.59	0.54-0.63	0.014	0.06	MgAl <sub>2</sub> O <sub>4</sub> , AlK $\alpha$ , TAP
Cr <sub>2</sub> O <sub>3</sub>	< D.L.	0.00-0.03	0.000	0.10	Synthetic MgCr <sub>2</sub> O <sub>4</sub> , CrK $\alpha$ , LIF
FeO	0.64	0.56-0.71	0.010	0.12	Fe <sub>2</sub> O <sub>3</sub> , FeK $\alpha$ , LIF
MnO	< D.L.	0.04-0.05	0.000	0.11	Synthetic MnSiO <sub>3</sub> , MnK $\alpha$ , LIF
MgO	0.44	0.40-0.48	0.013	0.06	Synthetic MgCr <sub>2</sub> O <sub>4</sub> , MgK $\alpha$ , TAP
CaO	46.52	46.43-46.62	0.976	0.07	CaMgSi <sub>2</sub> O <sub>6</sub> , CaK $\alpha$ , PET
NiO	< D.L.	0.00-0.04	0.000	0.15	Synthetic Ni <sub>2</sub> SiO <sub>4</sub> , NiK $\alpha$ , LIF
<b>Total</b>	<b>98.88</b>	<b>98.77-98.96</b>	<b>2.003</b>	<b>0.11</b>	

The data were averaged for three-spot analysis because of the very small size of the crystal. The cation ratios (calculated on the basis of three oxygen atoms) were calculated from the average concentrations listed in the second column. Na and K were not analysed. Ave, average; Stand. Dev., standard deviation; < D.L., lower than the detection limits (see Methods).

Extended Data Table 2 |  $d$  spacings, their corresponding relative intensities  $I$  (with respect to the most intense peak, at  $l=100$ ), and  $h, k, l$  indices for Ca-Pv, as obtained by single-crystal X-ray micro-diffraction

	CaSiO <sub>3</sub> (this study)		CaSiO <sub>3</sub> [23]	
$I$	$d_{\text{meas.}}$	$hkl$	$d$	$hkl$
24	3.81	1 1 0	3.831	1 1 0
<b>100</b>	<b>2.71</b>	<b>0 2 0</b>	<b>2.723</b>	<b>0 2 0</b>
<b>78</b>	<b>2.70</b>	<b>1 1 2</b>	<b>2.707</b>	<b>1 1 2</b>
<b>84</b>	<b>2.69</b>	<b>2 0 0</b>	<b>2.694</b>	<b>2 0 0</b>
25	2.20	2 0 2	2.203	2 0 2
<b>96</b>	<b>1.91</b>	<b>2 2 0</b>	<b>1.915</b>	<b>2 2 0</b>
43	1.57	1 3 2	1.569	1 3 2
34	1.56	3 1 2	1.558	3 1 2
37	1.36	2 2 4	1.354	2 2 4

The data are compared with those for a reference CaTiO<sub>3</sub> sample<sup>34</sup>.

**Extended Data Table 3 | Carbon isotopic composition ( $\delta^{13}\text{C}$ , in parts per thousand) and relative uncertainty for the host diamond enclosing the Ca-Pv inclusion**

Locations	$\delta^{13}\text{C}$	$2\sigma$
1	-2.3	0.5
2	-4.6	0.5
3	-2.4	0.5
4	-3.9	0.5
5	-3.9	0.4

The locations 1–5 refer to the positions noted in Fig. 1.



# Social norm complexity and past reputations in the evolution of cooperation

Fernando P. Santos<sup>1,2</sup>, Francisco C. Santos<sup>1,2</sup> & Jorge M. Pacheco<sup>2,3,4</sup>

**Indirect reciprocity is the most elaborate and cognitively demanding<sup>1</sup> of all known cooperation mechanisms<sup>2</sup>, and is the most specifically human<sup>1,3</sup> because it involves reputation and status. By helping someone, individuals may increase their reputation, which may change the predisposition of others to help them in future. The revision of an individual's reputation depends on the social norms that establish what characterizes a good or bad action and thus provide a basis for morality<sup>3</sup>. Norms based on indirect reciprocity are often sufficiently complex that an individual's ability to follow subjective rules becomes important<sup>4–6</sup>, even in models that disregard the past reputations of individuals, and reduce reputations to either 'good' or 'bad' and actions to binary decisions<sup>7,8</sup>. Here we include past reputations in such a model and identify the key pattern in the associated norms that promotes cooperation. Of the norms that comply with this pattern, the one that leads to maximal cooperation (greater than 90 per cent) with minimum complexity does not discriminate on the basis of past reputation; the relative performance of this norm is particularly evident when we consider a 'complexity cost' in the decision process. This combination of high cooperation and low complexity suggests that simple moral principles can elicit cooperation even in complex environments.**

Under indirect reciprocity, an individual expects a return not from someone whom they have helped directly but from a third party. Helping (or not helping) the 'right' individuals can increase the chance of being helped by someone else at a later stage<sup>9,10</sup>. Ohtsuki and Iwasa<sup>7,8,11</sup> defined a binary world in which an individual's reputation can be either 'good' or 'bad'. Even in such a simple world, an arbitrarily large set of associated social norms can be used to classify decisions made in a donation game. In each instance of this donation game, involving a 'donor' and a 'recipient', the donor may either cooperate, helping the recipient at a cost  $c$  to themselves while conferring a benefit  $b$  to the recipient (with  $b > c$ ), or defect (not providing help), in which case neither player incurs any costs or distributes any benefits. Everyone in the population uses the same social norm to assign public reputations to individuals. This reputation is attributed (errors aside; see Methods) and disseminated<sup>12–14</sup> by a bystander who witnesses a pairwise interaction. In this context, if all that matters for assigning a new reputation to the donor is their action towards the recipient<sup>10</sup>, then we have a first-order norm. If the current reputation of the recipient matters as well as the action of the donor, then we obtain a second-order norm. A third-order norm additionally includes the current reputation of the donor.

Most norms studied so far reach up to third order (see ref. 15 for an exception) and therefore rely, at most, on the action of the donor and on the current reputations of both the donor and the recipient. For a norm of a given order, the information used by an observer to assign a new reputation is the same information that a donor may use to decide how to act towards a recipient. Consequently, studies of indirect reciprocity involving norms of increasing order typically

use behavioural strategies (often designated action rules) and strategy spaces that also increase (exponentially with order). For this reason, a combination of a norm and a strategy that promotes cooperation in the space of  $n$ th-order norms does not necessarily perform equally well in a space of higher-order norms because the availability of more complex behaviours (together with those for lower-order norms) often has non-trivial effects on cooperation<sup>16</sup>. Furthermore, the performance of a complex social norm can be constrained by an individual's ability to follow complex subjective rules<sup>4–6</sup>. This raises two fundamental questions: (1) whether the moral principles that underlie successful strategies and norms in the space of third-order norms remain valid within a larger space, and if so which ones; and (2) how the cognitive skills associated with social norms and strategies impair individuals' performance. Using the donation game and binary reputations we answer these questions by investigating the cooperative capacity of social norms in a space that encompasses norms of up to fourth order and that span a wide range of cognitive complexities<sup>4,17,18</sup>. Increasing the number of possibilities to consider when assigning a good or a bad reputation to individuals enables us to identify the key pattern of social norms that provides the necessary conditions for promoting cooperation.

Fourth-order norms additionally incorporate (on top of the features of third-order norms) the previous reputation of the recipient, requiring individuals with increased memory capabilities and that are therefore able to enact more elaborate behaviours. We encode norms up to fourth order and corresponding strategies as 16- and 8-bit tuples, respectively; consequently, there are  $2^{16}$  different norms and  $2^8$  different strategies that individuals may use when playing the donation game described above (see Methods for details). Furthermore, we define the complexity of a norm using the index  $\kappa$ , which describes the number of literals (that is, the logic variables and their complements) in the shortest logical expression that can define the norm (see Methods). This index has been used previously to describe an individual's ability to learn a concept<sup>4,17</sup>. Here, the simplest norm has  $\kappa = 0$  and the most complex norm has  $\kappa = 32$ . In Fig. 1 we illustrate norms of different orders and complexities, providing intuitive representations of the raw information in Supplementary Table 4. Norms of the same order may have different complexities, as demonstrated for second-order norms in Fig. 1: different reputation tables (corresponding to different norms) translate to different numbers of literals in the corresponding minimal logical expressions. Moreover, similarly to norms, strategies also exhibit an intrinsic complexity ( $\kappa_s$ ) that can influence their adoption. Equipped with these tools, we investigate which norms promote the emergence of cooperation. In Methods, we describe computer simulations of the evolutionary dynamics, in which individuals in a population, each starting with a random strategy, play the donation game with their peers. Throughout the game, the players change strategies via social learning<sup>19</sup>, whereby strategies with higher fitness are adopted more frequently<sup>20</sup>. The simulations return the cooperation index  $\eta$ , a real number between 0 and 1 that

<sup>1</sup>INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016 Porto Salvo, Portugal. <sup>2</sup>ATP-group, 2744-016 Porto Salvo, Portugal. <sup>3</sup>Centro de Biologia Molecular Ambiental, Universidade do Minho, 4710-057 Braga, Portugal. <sup>4</sup>Departamento de Matemática e Aplicações, Universidade do Minho, 4710-057 Braga, Portugal.

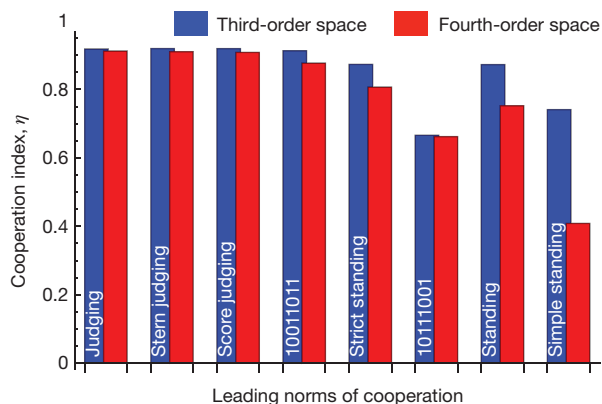
	Image score	Simple standing	Stern judging	Judging	Judging past																																																																																
Recipient ( $R_A, R_P$ )	<table><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr></table>	G	G	B	B	G	G	B	B	G	G	B	B	G	G	B	B	<table><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr><tr><td>G</td><td>G</td><td>G</td><td>G</td></tr><tr><td>G</td><td>G</td><td>G</td><td>G</td></tr></table>	G	G	B	B	G	G	B	B	G	G	G	G	G	G	G	G	<table><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr><tr><td>G</td><td>B</td><td>B</td><td>B</td></tr><tr><td>B</td><td>B</td><td>G</td><td>G</td></tr><tr><td>B</td><td>B</td><td>G</td><td>G</td></tr></table>	G	G	B	B	G	B	B	B	B	B	G	G	B	B	G	G	<table><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr><tr><td>B</td><td>B</td><td>B</td><td>G</td></tr><tr><td>B</td><td>B</td><td>B</td><td>G</td></tr></table>	G	G	B	B	G	G	B	B	B	B	B	G	B	B	B	G	<table><tr><td>G</td><td>G</td><td>B</td><td>B</td></tr><tr><td>G</td><td>B</td><td>B</td><td>B</td></tr><tr><td>B</td><td>B</td><td>B</td><td>G</td></tr><tr><td>B</td><td>B</td><td>G</td><td>G</td></tr></table>	G	G	B	B	G	B	B	B	B	B	B	G	B	B	G	G
	G	G	B	B																																																																																	
	G	G	B	B																																																																																	
	G	G	B	B																																																																																	
G	G	B	B																																																																																		
G	G	B	B																																																																																		
G	G	B	B																																																																																		
G	G	G	G																																																																																		
G	G	G	G																																																																																		
G	G	B	B																																																																																		
G	B	B	B																																																																																		
B	B	G	G																																																																																		
B	B	G	G																																																																																		
G	G	B	B																																																																																		
G	G	B	B																																																																																		
B	B	B	G																																																																																		
B	B	B	G																																																																																		
G	G	B	B																																																																																		
G	B	B	B																																																																																		
B	B	B	G																																																																																		
B	B	G	G																																																																																		
	( $R_D, A$ ) Donor	( $R_D, A$ ) Donor	( $R_D, A$ ) Donor	( $R_D, A$ ) Donor	( $R_D, A$ ) Donor																																																																																
Norm order	1	2	2	3	4																																																																																
Minimal DNF	$A$	$A \vee \overline{R_A}$	$R_A A \vee \overline{R_A} \overline{A}$	$R_A A \vee R_D \overline{R_A} \overline{A}$	$R_A A \vee R_D \overline{R_A} \overline{A} \vee R_P \overline{R_A} \overline{A}$																																																																																
Complexity, $\kappa$	1 = 1	1 + 1 = 2	2 + 2 = 4	2 + 3 = 5	2 + 3 + 3 = 8																																																																																

**Figure 1 | Norm complexity.** A norm is represented by a ‘reputation table’. Each entry in each table indicates the new reputation of the donor (good, G; bad, B), assigned on the basis of their current reputation ( $R_D \in \{G, B\}$ ), their action ( $A \in \{C, D\}$ , where C denotes cooperation and D defection), and the current ( $R_A \in \{G, B\}$ ) and past ( $R_P \in \{G, B\}$ ) reputations of the recipient. Rows are ordered, from top to bottom, as (G,G), (G,B), (B,B), (B,G) and columns are ordered, from left to right, as (G,C), (B,C), (B,D), (G,D). The complexity  $\kappa$  is determined by counting the number of literals

of the shortest logical expression (the minimal disjunctive normal form (DNF), where  $A$  denotes  $A = C$  and  $\bar{A}$  denotes the complement ( $\bar{A} = D$ ), and similarly  $R_{A,D}$  and  $\bar{R}_{A,D}$  denote G and B; see Methods) that can be used to prescribe a donor reputation of ‘G’. Alternatively,  $\kappa$  can be determined by counting the number of blocks of  $2^k$  ‘G’s<sup>30</sup> (where  $k$  is chosen to be as large as possible and blocks can overlap; see coloured squares and rectangles): each block of  $2^k$  ‘G’s increases  $\kappa$  by  $4 - k$  (starting from  $\kappa = 0$ ). See Supplementary Information for further details.

describes the average number of interactions that lead to donations as a fraction of the total number of interactions observed in a population that evolves under a given social norm.

In Fig. 2 we compare  $\eta$  for the leading eight norms shown<sup>7,8</sup> to stabilize cooperation (in the sense discussed in Supplementary Information, section 1.4) under indirect reciprocity at third order, in the space of third-order (blue bars) and fourth-order (red bars) norms. The results show that when more elaborate strategies become possible (when up to fourth-order norms are considered) only a subset of the leading eight norms still fosters similar levels of cooperation as in the third order space. Overall, about 0.2% of the  $2^{16}$  norms in fourth-order space lead to  $\eta > 0.9$ , compared to about 2% of the  $2^8$  norms in third-order space (Extended Data Fig. 1). Many ‘new’ fourth-order norms (that is, those that cannot be represented in lower-order spaces) foster high levels of cooperation. Of the leading two second-order norms<sup>21,22</sup> (stern judging and simple standing; see Supplementary Information for details), only stern judging remains highly cooperative in fourth-order space.

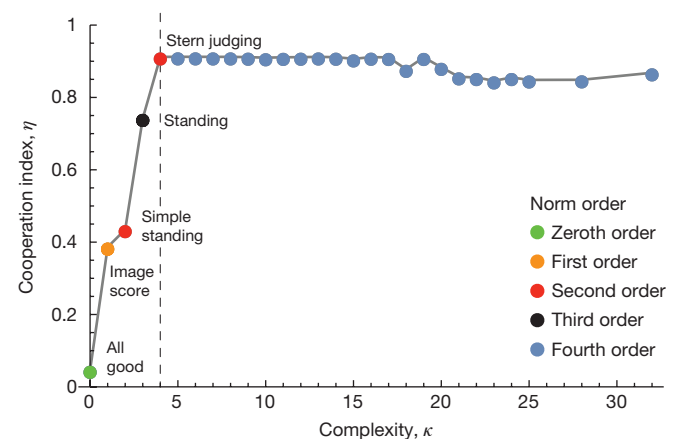


**Figure 2 | Cooperation index of leading norms.** When the space of the norms (and strategies) is extended from third-order (blue bars) to fourth-order (red bars), some of the leading eight norms of cooperation<sup>8</sup> (in third-order space)—and particularly simple standing (which, together with stern judging, make up the leading two norms in second-order space<sup>21</sup>)—no longer promote cooperation. See Extended Data Fig. 1 for results involving all norms. The model parameters used (see Methods for definitions) are  $Z = 50$ ,  $\varepsilon = \alpha = \chi = 0.01$ ,  $\mu = 1/Z$ ,  $b = 5$ ,  $c = 1$  and  $\gamma = 0$ . The results are qualitatively insensitive to the ratio  $b/c$ , to the population size, to any errors in assessment or assignments made by individuals and to different mutation schemes (see Methods and Extended Data Figs 4, 5). See Fig. 1 and Supplementary Table 4 for definitions and characterization of norms; unnamed norms are defined by their binary representation in third-order space (see Methods).

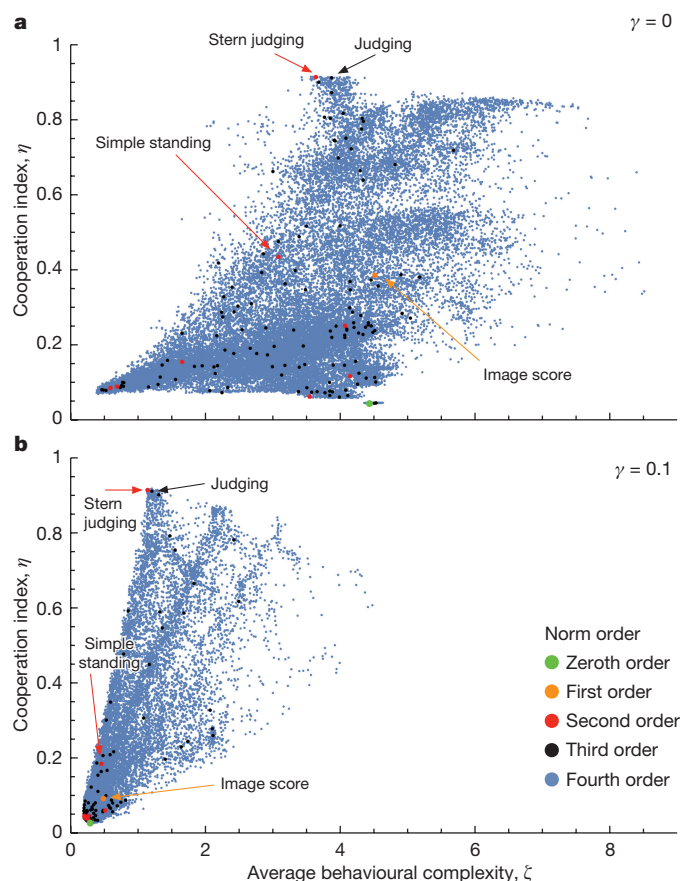
This norm can be stated as: “help good people and refuse help otherwise, and we shall be nice to you; otherwise, you will be punished.”<sup>23</sup>

Next, we investigate the role of norm complexity in promoting cooperation by plotting the cooperation level ( $\eta$ ) of the norm that leads to maximum cooperation for a given complexity ( $\kappa$ ). Figure 3 demonstrates that the highest values of  $\eta$  are attained by norms with complexities as low as  $\kappa = 4$ . The same happens even when individuals incur a complexity cost  $c_c = \gamma \kappa_s$  when using a strategy of complexity  $\kappa_s$  (where  $\gamma$  is a real constant; see Extended Data Figs 2 and 3 and Supplementary Information for details; we also demonstrate that these results remain valid when the past reputation of the donor instead of the recipient is used in defining fourth-order norms).

Figure 3 demonstrates that for  $\kappa > 4$  only fourth-order norms maximize  $\eta$ , despite the fact that the complexity of norms of the same order can vary substantially (see Fig. 1). Consequently, taking complexity into account opens up new questions regarding the features that make fourth-order norms successful, and the features of the third- and



**Figure 3 | Cooperation index versus norm complexity.** Maximal levels of cooperation ( $\eta > 0.9$ ) are attained under the simple norm stern judging ( $\kappa = 4$ ). More complex norms ( $\kappa > 4$ ) do not lead to higher levels of cooperation. Some well-known norms that maximize  $\eta$  for a given  $\kappa$  are identified. In Extended Data Fig. 2 we show the dependence of  $\eta$  on  $\kappa$  when a complexity cost is imposed on strategies and the past reputation of the donor is considered instead of that of the recipient. The model parameters used (see Methods for definitions) are  $Z = 50$ ,  $\varepsilon = \alpha = \chi = 0.01$ ,  $\mu = 1/Z$ ,  $b = 5$ ,  $c = 1$  and  $\gamma = 0$ . See Extended Data Figs 4 and 5 for robustness analysis. See Fig. 1 and Supplementary Table 4 for definitions and characterization of norms.



**Figure 4 | Average behavioural complexity.** Norms induce different levels of strategic complexity  $\kappa_s$  in a population. Because the simplest strategies ignore reputations, and are thus unable to secure cooperation, we expect high levels of cooperation to require some average behavioural complexity  $\zeta$ . **a**, We find that stern judging and judging lead to maximum values of  $\eta$  at relatively low values of  $\zeta$ . **b**, This finding is emphasized if we consider a strategic complexity cost  $c_c = \gamma\kappa_s$  with  $\gamma \neq 0$  (see Extended Data Fig. 2). The model parameters used, detailed in Methods, are  $Z = 50$ ,  $\varepsilon = \alpha = \chi = 0.01$ ,  $\mu = 1/Z$ ,  $b = 5$  and  $c = 1$ . See Fig. 1 for definitions of judging, stern judging, simple standing and image score.

second-order norms that ensure (or not) their capacity to sustain cooperation in the more complex fourth-order space.

To address these questions, we conducted an exhaustive search in the space of fourth-order norms and identified (for a specific set of model parameters) a recurrent pattern common to the fourth-order norms that promote cooperation (see Supplementary Tables 1 and 2). This pattern states that the bystander assigns a ‘good’ label to donors that either (i) cooperate with enduring good individuals or (ii) are already good and defect against enduring bad individuals, and assigns a ‘bad’ label to those who act otherwise in these contexts (that is, who defect against enduring good individuals or who are good but cooperate with enduring bad individuals). Here, enduring individuals are those who retain the same good or bad label in the present and in the past. The pattern can therefore be summarized by the following rule: “donors become good (bad) if they help (refuse to help) an enduring good individual; they maintain (lose) their good label if they refuse to help (help) an enduring bad individual.”

This rule has immediate implications at lower orders. Only four of the leading eight norms<sup>8</sup> in third-order space comply with this fourth-order rule—those that promote the highest levels of cooperation (Fig. 2). Not surprisingly (see Fig. 3), stern judging is the only one of the leading two<sup>21</sup> norms in second-order space that complies (simple standing violates the rule by prescribing a good reputation whenever a player helps an enduring bad individual).

In Fig. 3 we show that stern judging leads to a maximal value of  $\eta$  ( $\eta > 0.9$ ), while having a  $\kappa$  value less than that of any third- or fourth-order social norm that leads to comparable values of  $\eta$  (see also Extended Data Fig. 2). Furthermore, strategies that prevail under stern judging are remarkably simple. We demonstrate this by first computing the complexity  $\kappa_s$  of the prevalent strategies under each norm. Subsequently, we compute the (norm-dependent) fraction of time that each individual spends adopting each strategy and calculate the weighted average complexity of the strategies used, which we designate by the average behavioural complexity ( $\zeta$ ). In Fig. 4 we depict all norms in fourth-order space by plotting  $\eta$  as a function of  $\zeta$ . Stern judging (a second-order norm), judging and score judging (third-order norms; see Supplementary Table 4) lead to high  $\eta$  using strategies with low  $\zeta$  (Fig. 4a)—a feature that is maintained in the presence of a complexity cost  $c_c = \gamma\kappa_s$  (Fig. 4b).

Our results show that cooperation under indirect reciprocity can emerge even when the cognitive capacity of individuals is limited. In this context, it becomes clear why stern judging proves to be so robust, remaining the most successful norm (in terms of the combination of high cooperation and low complexity) in all norm spaces studied even when considering populations of different sizes (from small-scale societies to large communities of individuals<sup>22</sup>). It is the norm of lowest order and complexity that is compatible with the pattern described here, requiring little cognitive skill both in assigning reputations and in inducing behaviours that lead to high levels of cooperation. It is therefore not surprising that the fingerprint of stern judging is present in the moral judgment of toddlers (as young as five months old<sup>24</sup>), who show a preference not only for individuals who helped others, but also for individuals who harmed those who hindered others<sup>25</sup>.

The modelling approach used here can also be informative when designing pervasive reputation systems<sup>26</sup>, in which optimality should be combined with simplicity. Game-theoretical models have been used to study reputation systems in the context of trading platforms, crowdsourcing markets and peer-to-peer systems<sup>27–29</sup>. It has been shown that very simple and intuitive social norms may suffice to promote cooperation<sup>28</sup> and that publicizing a detailed account of a seller’s feedback history—as compared with only the most recent rating—does not improve cooperation in online trading platforms<sup>27</sup>. Both of these features—simplicity and the irrelevance of history—bear similarity to the results presented here, despite the fact that our model would need to be modified to be applicable to reputation systems in online platforms.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 5 September 2017; accepted 15 January 2018.**

- Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
- Rand, D. G. & Nowak, M. A. Human cooperation. *Trends Cogn. Sci.* **17**, 413–425 (2013).
- Alexander, R. D. *The Biology of Moral Systems* (Transaction Publishers, 1987).
- Feldman, J. Minimization of Boolean complexity in human concept learning. *Nature* **407**, 630–633 (2000).
- Chater, N. & Vitányi, P. Simplicity: a unifying principle in cognitive science? *Trends Cogn. Sci.* **7**, 19–22 (2003).
- Feldman, J. The simplicity principle in perception and cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **7**, 330–340 (2016).
- Ohtsuki, H. & Iwasa, Y. How should we define goodness?—reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231**, 107–120 (2004).
- Ohtsuki, H. & Iwasa, Y. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435–444 (2006).
- Brandt, H. & Sigmund, K. The logic of reprobation: assessment and action rules for indirect reciprocity. *J. Theor. Biol.* **231**, 475–486 (2004).
- Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
- Ohtsuki, H., Iwasa, Y. & Nowak, M. A. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* **457**, 79–82 (2009).
- Dunbar, R. *Grooming, Gossip, and the Evolution of Language* (Harvard Univ. Press, 1998).

13. Sommerfeld, R. D., Krambeck, H.-J., Semmann, D. & Milinski, M. Gossip as an alternative for direct observation in games of indirect reciprocity. *Proc. Natl Acad. Sci. USA* **104**, 17435–17440 (2007).
14. Skyrms, B. *Signals: Evolution, Learning and Information* (Oxford Univ. Press, 2010).
15. Kandori, M. Social norms and community enforcement. *Rev. Econ. Stud.* **59**, 63–80 (1992).
16. Stewart, A. J., Parsons, T. L. & Plotkin, J. B. Evolutionary consequences of behavioral diversity. *Proc. Natl Acad. Sci. USA* **113**, E7003–E7009 (2016).
17. Wegener, I. & Teubner, B. *The Complexity of Boolean Functions* Vol. 1 (B. G. Teubner, 1987).
18. McCluskey, E. J. Minimization of Boolean functions. *Bell Labs Tech. J.* **35**, 1417–1444 (1956).
19. Rendell, L. *et al.* Why copy others? Insights from the social learning strategies tournament. *Science* **328**, 208–213 (2010).
20. Sigmund, K. *The Calculus of Selfishness* (Princeton Univ. Press, 2010).
21. Ohtsuki, H. & Iwasa, Y. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *J. Theor. Biol.* **244**, 518–531 (2007).
22. Santos, F. P., Santos, F. C. & Pacheco, J. M. Social norms of cooperation in small-scale societies. *PLOS Comput. Biol.* **12**, e1004709 (2016).
23. Pacheco, J. M., Santos, F. C. & Chalub, F. A. C. Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. *PLOS Comput. Biol.* **2**, e178 (2006).
24. Hamlin, J. K. Moral judgment and action in preverbal infants and toddlers evidence for an innate moral core. *Curr. Dir. Psychol. Sci.* **22**, 186–193 (2013).
25. Hamlin, J. K., Wynn, K., Bloom, P. & Mahajan, N. How infants and toddlers react to antisocial others. *Proc. Natl Acad. Sci. USA* **108**, 19931–19936 (2011).
26. Resnick, P., Kuwabara, K., Zeckhauser, R. & Friedman, E. Reputation systems. *Commun. ACM* **43**, 45–48 (2000).
27. Dellarocas, C. Reputation mechanism design in online trading environments with pure moral hazard. *Inform. Syst. Res.* **16**, 209–230 (2005).
28. Ho, C.-J., Zhang, Y., Vaughan, J. & Van Der Schaar, M. *Towards Social Norm Design for Crowdsourcing Markets*. Report No. WS-12-08 (AAAI, 2012).
29. Zhang, Y. & van der Schaar, M. Peer-to-peer multimedia sharing based on social norms. *Signal. Process. Image Commun.* **27**, 383–400 (2012).
30. Karnaugh, M. The map method for synthesis of combinational logic circuits. *Trans. AIEE Part I* **72**, 593–599 (1953).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported by Fundação para a Ciência e Tecnologia (FCT) through grants SFRH/BD/94736/2013, PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014, UID/BIA/04050/2013 and UID/CEC/50021/2013. We are grateful to A. P. Francisco and M. Janota for comments.

**Author Contributions** F.P.S., F.C.S. and J.M.P. conceived the project. F.P.S. performed the mathematical and numerical analysis. F.P.S., F.C.S. and J.M.P. analysed the results and wrote the paper. All authors contributed to all other aspects of the project.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.M.P. ([jmpacheco@math.uminho.pt](mailto:jmpacheco@math.uminho.pt)).

**Reviewer Information** *Nature* thanks C. Efferson, E. Fehr, G. Szabó and A. Tavoni for their contribution to the peer review of this work.



## METHODS

Here we summarize the model and mathematical methods; further details are provided in Supplementary Information.

**Actions conditional on reputations.** The action of the donor in each interaction depends on the current reputation of the donor ( $R_D$ ) and the recipient ( $R_A$ ), together with the past reputation of the recipient ( $R_P$ ). Assuming binary reputations ( $1 = \text{'good'} = G$  or  $0 = \text{'bad'} = B$ ), the strategy used by each player is an 8-bit string that prescribes an action ( $1 = \text{'cooperate'} = C$  or  $0 = \text{'defect'} = D$ ) on the basis of the aforementioned reputations. We extend previously used notation<sup>7,8,21</sup> to denote each strategy by a tuple  $P = (p_0, p_1, p_2, p_3, p_4, p_5, p_6, p_7)$ , in which  $p_i \in \{0, 1\}$  denotes the action of the donor for each of the possible combinations of reputations  $R$  in the order  $R_P, R_D, R_A$  (that is, with  $R_P, R_A$  being the most (least) significant bit when defining a position within a strategy), and with  $R_i = 1$  considered before  $R_i = 0$  (that is, for example,  $p_0$  corresponds to  $R_P = R_D = R_A = G = 1$  and  $p_7$  to  $R_P = R_D = R_A = B = 0$ ); this yields  $2^8$  different strategies. We consider execution errors ( $\varepsilon$ ) that represent the inability of individuals to act in the way that their strategy dictates<sup>31</sup>. It is common practice to consider errors in the form of 'failed intended cooperation'<sup>21,32</sup> due, for instance, to an individual's lack of resources, time or energy available to donate in their role as donor<sup>33</sup>. Our results remain valid even if the execution errors additionally induce defectors to involuntarily cooperate.

**Social norms.** We consider that the new reputation of an acting individual follows a norm that can be written as a tuple  $d = (d_0, d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{14}, d_{15})$ , in which  $d_i \in \{0, 1\}$  denotes the new reputation assigned to the donor for each of the possible combinations of action  $A$  and reputations  $R$  in the order  $R_P, R_D, R_A, A$  (that is, with  $R_P, A$  being the most (least) significant bit when defining a position within a norm). For convenience, we use  $R_P, R_D$  and  $R_A$  both as the names of a reputation layer in a norm (see Extended Data Fig. 3 and Supplementary Table 3) and as a Boolean variable that can assume the values  $1 = G = R$  and  $0 = B = \bar{R}$ . Similarly,  $1 = C = A$  and  $0 = D = \bar{A}$ . As stated in the main text (see Fig. 1), there are  $2^{16}$  social norms up to fourth order. We consider assignment errors<sup>8</sup>  $\alpha$  that occur when the observer fails to assign the correct reputation. We assume that, once the reputation of an individual is assigned, it is widely disseminated throughout the population (for example through gossiping<sup>11–14</sup>), so that everyone shares the same opinion regarding the reputation of others. However, we include errors at the level of individuals, when retrieving the public reputation of others, which occur with a probability  $\chi$ : whenever these errors occur, an individual may perform the wrong action as a donor or assign the wrong (public) reputation as a bystander.

**Complexity.** Social norms and individual strategies can both be regarded as Boolean functions that determine: (1) when an individual has a good reputation ( $G$ ; social norms), or (2) when the appropriate action is to cooperate ( $C$ ; strategies or action rules). These functions take the Boolean inputs  $A$  (action of the donor is  $C$ ),  $R_A$  (current reputation of the recipient is  $G$ ),  $R_P$  (past reputation of the recipient or donor is  $G$ ) and  $R_D$  (current reputation of the donor is  $G$ ). For instance, the well-known second-order discriminator strategy whereby an individual cooperates with only those players who have a  $G$  reputation is given by  $P = (1, 0, 1, 0, 1, 0, 1, 0)$ , or by the Boolean function  $R_A$ . The fourth-order discriminator strategy, whereby an individual cooperates only if an opponent has a  $G$  reputation both in the present and in the past, can be written as  $P = (1, 0, 1, 0, 0, 0, 0, 0)$  or  $R_A \wedge R_P$ . In the context of social norms, the 'image score' norm  $d = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)$  corresponds to  $R_A$ , and the 'stern judging' norm  $d = (1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1)$  can be written as  $(R_A \wedge A) \vee (\bar{R}_A \wedge \bar{A})$ . The complexity of a norm or strategy ( $\kappa$  or  $\kappa_s$ ) is the length of the shortest Boolean formula (here in disjunctive normal form (DNF); that is, a sum of products) that is logically equivalent to the corresponding Boolean function<sup>4,17</sup>. This quantity is also known as the Boolean complexity<sup>4,17</sup>. To calculate the Boolean complexity of a norm or strategy, we generate and simplify the corresponding DNF and count the number of literals that it includes. We apply a standard algorithm to minimize Boolean functions (the Quine–McCluskey algorithm<sup>18</sup>), using the version implemented in Mathematica (Wolfram) through the function *BooleanMinimize*. This algorithm generates a DNF with a minimum number of literals but that is logically equivalent to the original (full) DNF—the minimal DNF (see Fig. 1). Here we focus on the minimal DNF representation of a logic expression. However, other representations could be devised in which, in some cases, there is departure from a minimal DNF and the number of literals is reduced slightly—such as by applying De Morgan's laws and/or the distributive law of Boolean algebra<sup>18</sup>. In fact, reaching a minimal Boolean function is a computational challenge<sup>34</sup>, and for this reason it is often calculated as an approximation<sup>4,35,36</sup>. By adopting a complexity measure based on the number of literals of a minimal DNF form, we provide an upper bound on the Boolean complexity of each social norm, while ensuring computational tractability and an easy generalization to norms of higher order. In Supplementary Information

we define (and provide an example of) the three-step process that we use to compute  $\kappa$  for any norm (and  $\kappa_s$  for any strategy).

In Fig. 1 we also provide an alternative visual method to determine  $\kappa$ . It relies on counting the number of different blocks of 'G's of size  $2^k$ , a method that is associated with so-called Karnaugh maps<sup>30</sup> (a graphical method for simplifying logic circuits): a size- $2^3$  block contributes 1 to the complexity; a size- $2^2$  block contributes 2; a size- $2^1$  block contributes 3; and a size- $2^0$  block contributes 4. In general, a  $2^k$ -size  $G$  block contributes  $4 - k$  to  $\kappa$ . Some rules apply when defining  $G$  blocks<sup>37</sup>: they must contain only  $G$  values, being formed by joining adjacent cells (diagonal links do not count); torus boundary conditions apply; and they must be the largest possible size. Importantly, the choice of row and column order in defining the reputation table in Fig. 1 is not arbitrary: the entries in two adjacent rows or columns must differ only by one bit.

It is also worth pointing out that Fig. 1 provides visual cues that show the symmetries of a reputation table that are associated with a norm of a given order; for example, for norms of order one, all of the entries of the left and right eight-entry blocks are identical. In all cases in Fig. 1, blocks of entries are delimited by solid lines: norms of second order have four blocks that each contain four identical entries; norms of third order have eight blocks that each contain two identical entries; and norms of fourth order have no such blocks in which multiple identical entries can be identified.

**Evolutionary dynamics.** In the computer simulations, evolution proceeds in discrete steps. At the beginning of one simulation (or run), each individual adopts one of the  $2^8$  (256) possible strategies, chosen using a uniform probability distribution (UPD). Individual reputations, both present and past, are also assigned using a UPD. Each simulation is executed for a large number  $g$  of generations. In each generation,  $Z$  individuals selected using a UPD revise their strategy. After selecting one of the  $Z$  individuals (say, individual  $X$ ), strategy revision can happen through mutation or imitation. Mutation<sup>38</sup> happens with probability  $\mu$ : a new strategy is adopted randomly (UPD) out of the 256 possible. This approach allows us to study the evolutionary robustness of strategies against the invasion of others<sup>39–42</sup> (see Supplementary Information). Alternatively, we consider a bit-wise (or local) mutation (see Extended Data Fig. 5), which leads to similar results. Imitation happens with probability  $1 - \mu$ : a new individual (say, individual  $Y$ , the role model) is selected randomly, and individual  $X$  is given the opportunity to update their strategy. The fitness of both individuals ( $F_X$  and  $F_Y$ ) is calculated as the average payoff earned in  $g = 2Z$  games played against individuals in the population selected randomly using a UPD. This number of games is adequate to obtain a clear assessment of the average payoff, given the number of strategies, and to account for the dynamic reputation assignment described below. After each game is played, a reputation update occurs according to the social norm and subject to the assessment ( $\alpha$ ) and private ( $\chi$ ) errors described above. Individual  $X$  adopts the strategy of individual  $Y$  with probability  $(1 + e^{F_X - F_Y})^{-1}$ —the so-called Fermi update or pairwise comparison rule<sup>43</sup>.

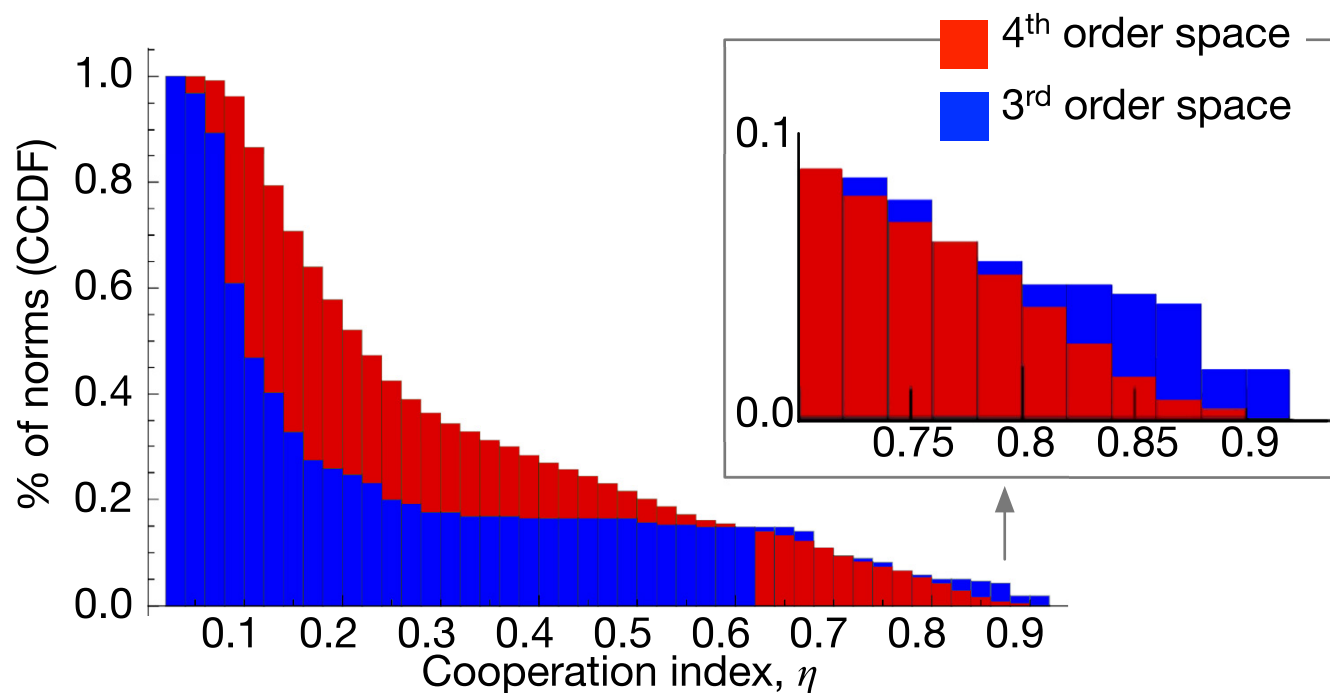
**Cooperation index.** The cooperation index  $\eta$  for a given social norm is computed as the fraction of cooperative acts that take place out of the total number of acts during the simulation time. Thus,  $\eta$  reflects both the dependence of strategy adoption on the relative frequency of strategies present in the population (frequency-dependent selection) and the evolution of reputations given the fixed social norm in the population. More details on the computer simulations are provided in Supplementary Information and in Extended Data Fig. 6. The full set of parameters explored is summarized in Supplementary Table 2.

**Code availability.** A comprehensive description of the standard algorithms that we implemented to compute the evolutionary dynamics of strategies is provided in Supplementary Information and Extended Data Fig. 6. Code that exemplifies the calculation of Boolean complexity is available at <https://doi.org/10.5281/zenodo.1041379>.

**Data availability.** The raw data generated, which were used to create Figs 2–4 and which support our conclusions, is available with the online version of the paper as Source Data.

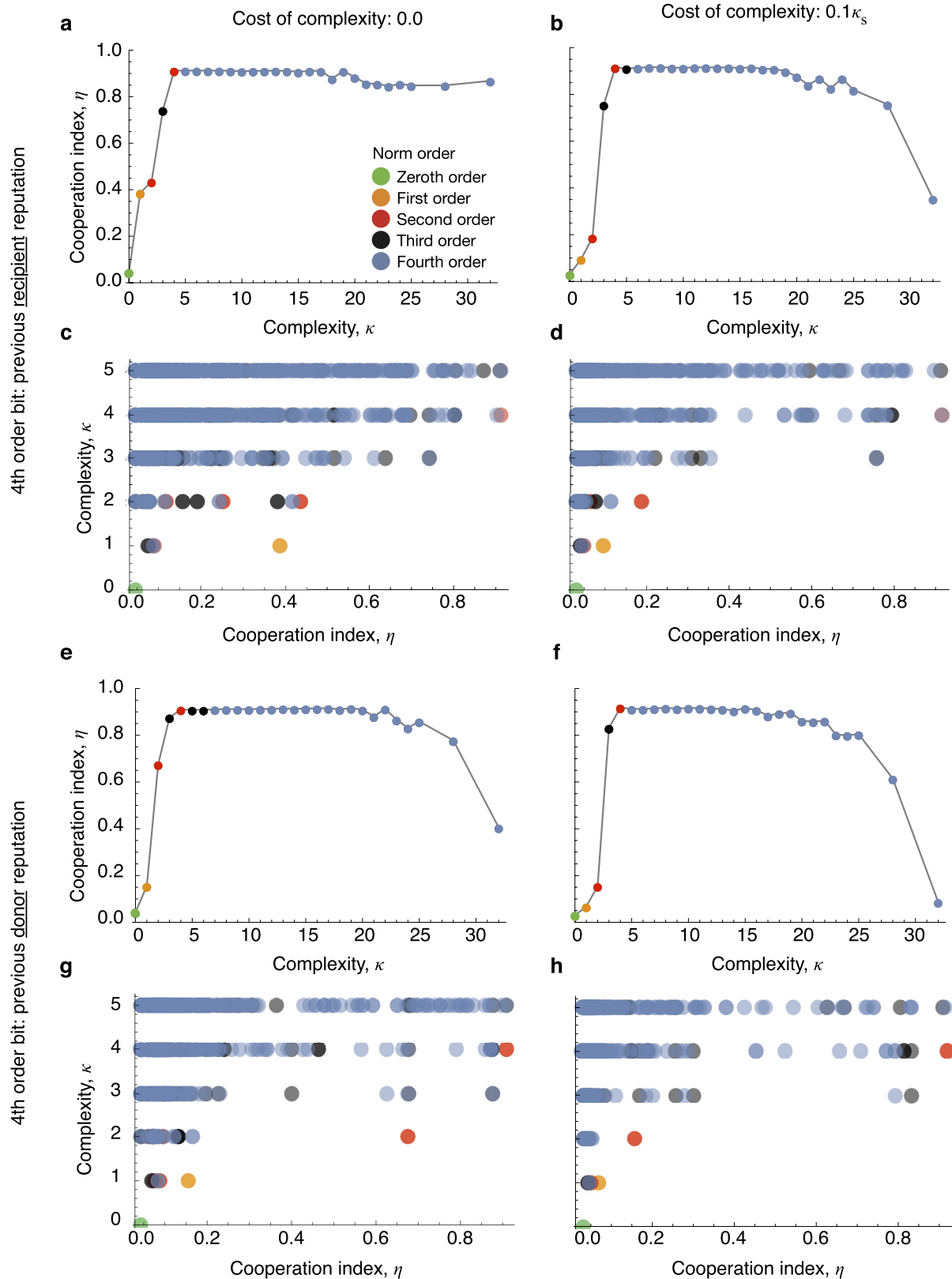
1. Fishman, M. A. Indirect reciprocity among imperfect individuals. *J. Theor. Biol.* **225**, 285–292 (2003).
2. Roberts, G. Evolution of direct and indirect reciprocity. *Proc. R. Soc. Lond. B* **275**, 173–179 (2008).
3. Sherratt, T. N. & Roberts, G. The importance of phenotypic defectors in stabilizing reciprocal altruism. *Behav. Ecol.* **12**, 313–317 (2001).
4. Umans, C. The minimum equivalent DNF problem and shortest implicants. *J. Comput. Syst. Sci.* **63**, 597–611 (2001).
5. Vigo, R. A note on the complexity of Boolean concepts. *J. Math. Psychol.* **50**, 501–510 (2006).
6. Feldman, J. The simplicity principle in human concept learning. *Curr. Dir. Psychol. Sci.* **12**, 227–232 (2003).

37. Null, L. & Lobur, J. *The Essentials of Computer Organization and Architecture* Ch. 3 (Jones & Bartlett Publishers, 2014).
38. Santos, F. P., Pacheco, J. M. & Santos, F. C. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci. Rep.* **6**, 37517 (2016).
39. Stewart, A. J. & Plotkin, J. B. From extortion to generosity, evolution in the iterated prisoner's dilemma. *Proc. Natl Acad. Sci. USA* **110**, 15348–15353 (2013).
40. Stewart, A. J. & Plotkin, J. B. Collapse of cooperation in evolving games. *Proc. Natl Acad. Sci. USA* **111**, 17558–17563 (2014).
41. Pinheiro, F. L., Vasconcelos, V. V., Santos, F. C. & Pacheco, J. M. Evolution of all-or-none strategies in repeated public goods dilemmas. *PLOS Comput. Biol.* **10**, e1003945 (2014).
42. Hilbe, C., Martinez-Vaquero, L. A., Chatterjee, K. & Nowak, M. A. Memory- $n$  strategies of direct reciprocity. *Proc. Natl Acad. Sci. USA* **114**, 4715–4720 (2017).
43. Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* **74**, 011909 (2006).



**Extended Data Figure 1 | Cooperation index of third- and fourth-order norms.** In the space of fourth-order norms (red bars), only a small fraction of norms (about 0.2% of  $2^{16}$ ) foster high levels of cooperation ( $\eta > 0.9$ ), as conveyed by the complementary cumulative distribution function (CCDF;

see inset for a close-up of the tail). In the space of third-order norms (blue bars), about 2% of norms (of a total of  $2^8$ ) promote high levels of cooperation ( $\eta > 0.9$ ). Other parameters:  $Z = 50$ ,  $\varepsilon = \alpha = \chi = 0.01$ ,  $\mu = 1/Z$ ,  $b = 5$ ,  $c = 1$ ,  $\gamma = 0$ .

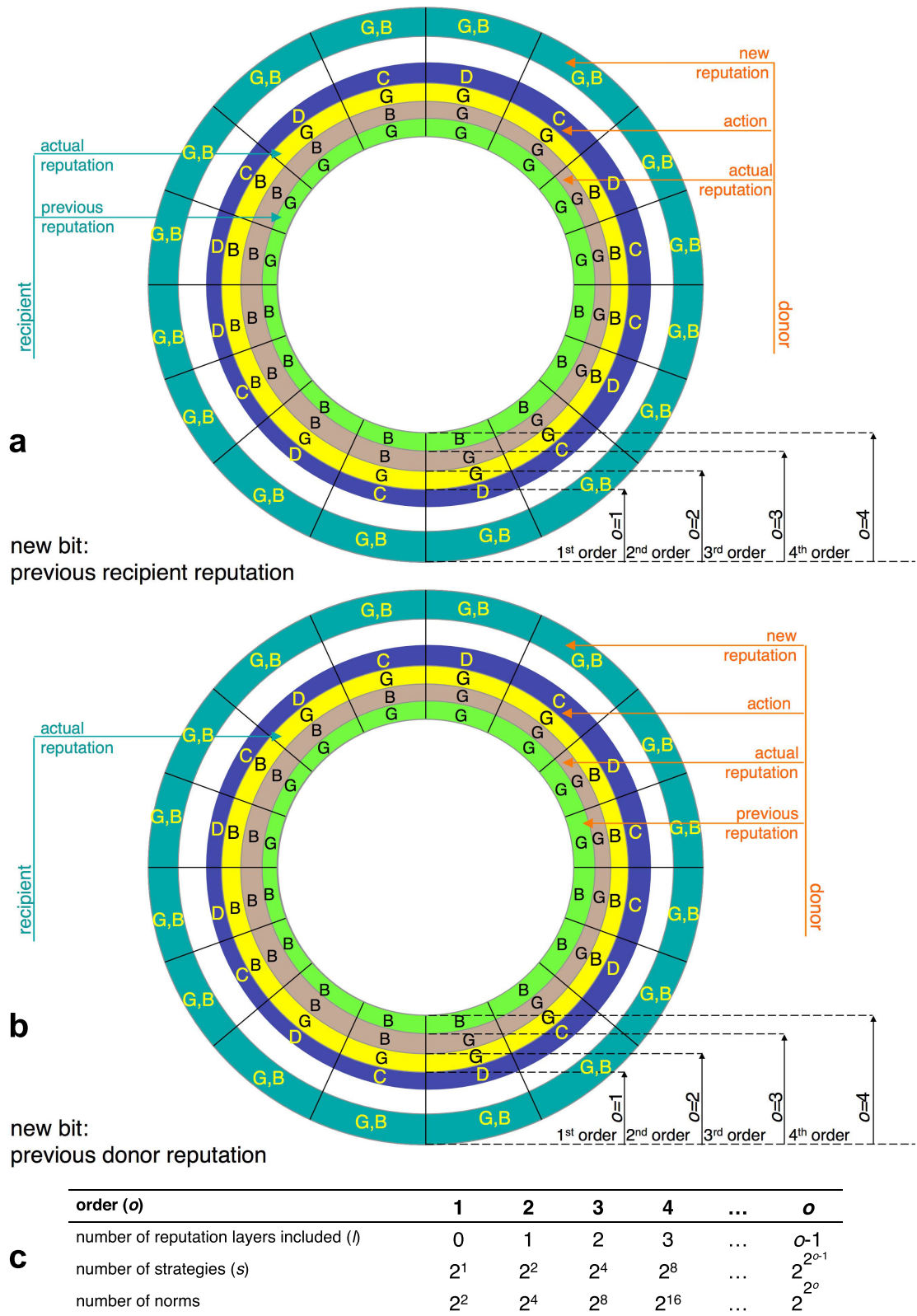


### Extended Data Figure 2 | The most cooperative norms.

**a, c, e, g,** Data from simulations in which individuals pay a complexity cost  $c_c$  proportional to the complexity  $\kappa_s$  of the strategy that they employ ( $c_c = c\kappa_s/10 = \gamma\kappa_s$ ). **b, d, f, h,** Data when no complexity cost is involved. Irrespective of whether the previous reputation of the recipient (**a–d**) or the donor (**e–h**) is used as the fourth consideration (as the fourth-order bit; see Extended Data Fig. 3), or whether there is a complexity cost involved, the highest levels of cooperation are already achieved for  $\kappa = 4$ .

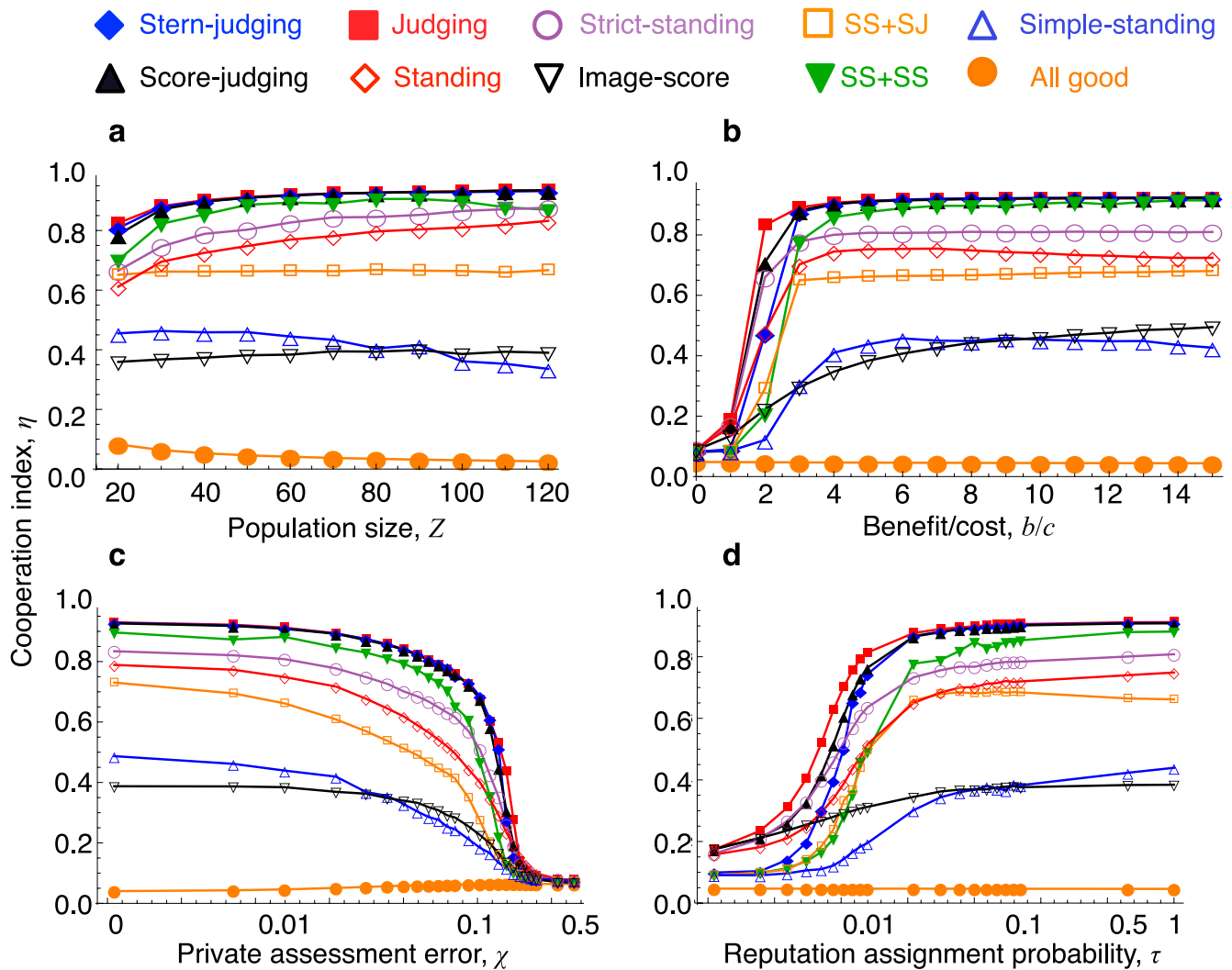
Moreover, when we plot norm performance (in terms of the cooperation index), separating norms according to their complexity  $\kappa$  (for  $\kappa \leq 5$ ; **c, d, g and h**) it becomes apparent that fourth-order norms are generally outperformed by lower order norms. Furthermore, paying a complexity cost is most detrimental to the more sophisticated fourth-order norms, which no longer promote cooperation under indirect reciprocity. Other parameters:  $Z = 50$ ,  $\varepsilon = \alpha = \chi = 0.01$ ,  $\mu = 1/Z$ ,  $b = 5$ ,  $c = 1$ .





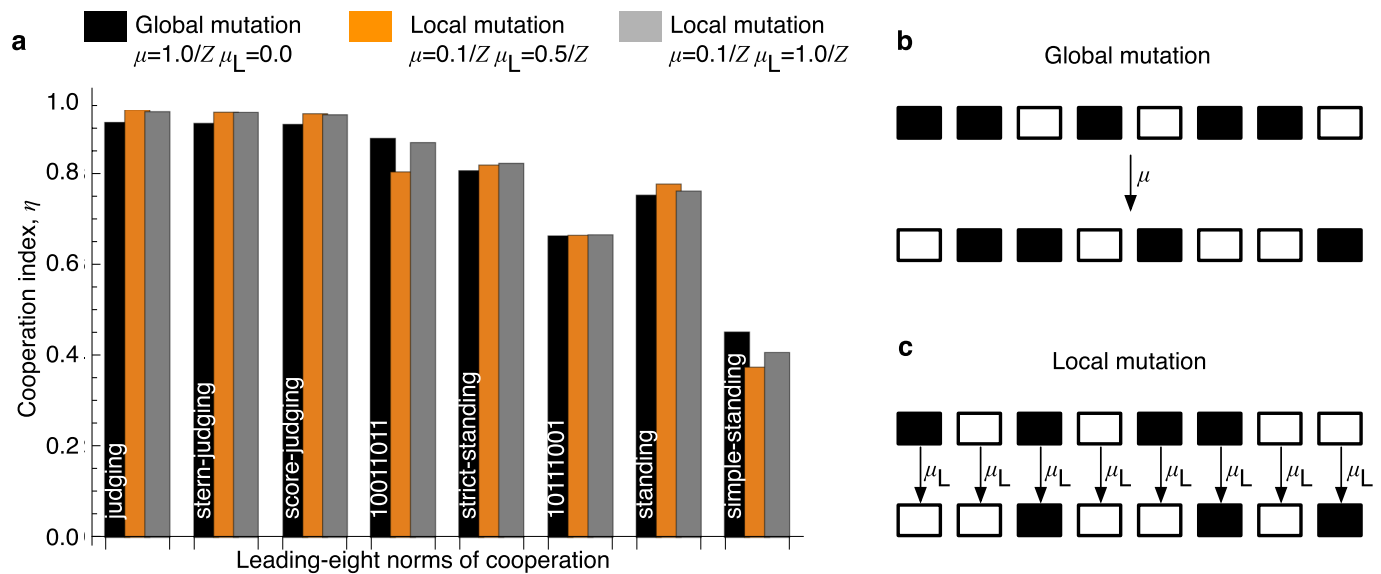
**Extended Data Figure 3 | Alternative ways of defining a social norm.** **a, b,** We consider norms that attribute a new reputation (outer ring) on the basis of (i) the action of the donor (first-order bit; blue ring); (ii) the current (actual) reputation of the receiver (second-order bit; yellow ring); (iii) the current (actual) reputation of the donor (third-order bit; pink ring); and (iv) the previous reputation of either the recipient (**a**) or the

donor (**b**) (fourth-order bit; green ring). In **a** and **b**, there are  $2^{16}$  norms in total. **c,** Number of bits (layers,  $l$ ) used for each norm order, and the corresponding number of possible strategies ( $s$ ) and norms. Because actions are taken using the same information used by a norm to attribute a new reputation, we consider  $2^8$  different strategies for norms up to fourth-order.



**Extended Data Figure 4 | Robustness of results to parameter variations.** A full list and detailed description of all model parameters is provided in Supplementary Information. **a–d**, Norm performance (in terms of the complexity index) as a function of population size  $Z$  (**a**), the benefit-to-cost ratio  $b/c$  in the donation game in which individuals interact (**b**), the private assessment error probability  $\chi$  (**c**) and the reputation assignment probability  $\tau$  (**d**). Here we use the previous reputation of the recipient as the fourth-order bit (as in the main text) and investigate, within the

space of fourth-order norms, the performance of the (second- and third-order) leading eight norms together with the (first-order) image score and (zeroth-order) all good norms. The norms 'ss' and 'sj' denote simple standing and stern judging; 'ss + sj' has the first 8 bits equal to the third-order representation of simple standing and the last 8 equal to the third-order representation of stern judging; and 'sj + ss' is defined similarly; see Supplementary Table 4 for details of these norms. Other parameters:  $Z = 50$ ,  $\varepsilon = \alpha = \chi = 0.01$ ,  $\mu = 1/Z$ ,  $b = 5$ ,  $c = 1$ ,  $\gamma = 0$ .



#### Extended Data Figure 5 | Global versus local mutation schemes.

**b**, We consider a mutation scheme in which a new strategy is adopted with probability  $\mu$  (drawn from a UPD) when a mutation occurs<sup>39–42</sup>.

**c**, Alternatively, we consider a local mutation (in each strategy), whereby

with probability  $\mu_L$  (drawn from a UPD) one bit changes. **a**, For the leading eight norms<sup>8</sup>, we find that the same conclusions are attained regardless of the mutation scheme considered. Other parameters:  $Z = 50$ ,  $\varepsilon = \alpha = \chi = 0.01$ ,  $b = 5$ ,  $c = 1$ ,  $\gamma = 0$ .

**Runs**: number of runs;  
**Gens**: number of generations;  
**Z**: population size;  
**P**: vector of all individual strategies;  
 $P_k$ : strategy of individual  $k$ ;  
**R**: vector of all individual public reputations;  
 $R^k$ : public reputation of individual  $k$ ;  
 $R_P^k$ : previous public reputation of individual  $k$ ;  
 $\mathcal{U}\{a, b\}$ : uniform distribution over integers between  $a$  and  $b$ ;  
 $Rand \sim \mathcal{U}(0, 1)$ : random value sampled following the standard uniform distribution;  
 $F_a$ : fitness of individual  $a$ ;  
 $\gamma$ : behavioural complexity cost;  
 $\kappa(P_k)$ : boolean complexity of strategy  $P_k$ ;  
 $d(A, R_A, R_D, R_P)$ : new reputation given social norm  $d$ , action  $A$ , recipient actual reputation  $R_A$ , donor actual reputation  $R_D$  and previous reputation  $R_P$ ;  
 $\Pi(x, y)$ : payoff to individual  $x$  after an interaction with  $y$  where both  $x$  and  $y$  may donate following their strategies ( $P_x$  and  $P_y$ ). A potential update of reputations ( $R^x$  and  $R^y$ ) occurs with probability  $\tau$ . This step takes into account the execution, assignment and private assessment errors.  
**Cooperate**  $\equiv 1$ ; **Defect**  $\equiv 0$ ;  
**Good**  $\equiv 1$ ; **Bad**  $\equiv 0$ ;

```

for  $r \leftarrow 1$  to  $Runs$  do
  for  $k \leftarrow 1$  to  $Z$  do
     $P_k \leftarrow X \sim \mathcal{U}\{0, 255\}$ 
     $R_P^k \leftarrow X \sim \mathcal{U}\{0, 1\}$ 
     $R^k \leftarrow X \sim \mathcal{U}\{0, 1\}$ 
  end
  for  $t \leftarrow 1$  to  $Gens$  do
     $a \leftarrow X \sim \mathcal{U}\{1, Z\}$ 
    if  $Rand < \mu$  then  $P_a \leftarrow X \sim \mathcal{U}\{0, 255\}$ ;
    else
       $b \leftarrow X \sim \mathcal{U}\{1, Z\}$ ,  $b \neq a$ ;
       $F_a \leftarrow 0$ ;
       $F_b \leftarrow 0$ ;
      for  $i \leftarrow 1$  to  $2Z$  do
         $c \leftarrow X \sim \mathcal{U}\{1, Z\}$ ,  $c \neq a$ ;
         $F_a \leftarrow F_a + \Pi(a, c) - \kappa(P_a)\gamma$ ;
        /* update  $R_P^a$ ,  $R_P^c$ ,  $R^a$ ,  $R^c$  */
         $c \leftarrow X \sim \mathcal{U}\{1, Z\}$ ,  $c \neq b$ ;
         $F_b \leftarrow F_b + \Pi(b, c) - \kappa(P_b)\gamma$ ;
        /* update  $R_P^b$ ,  $R_P^c$ ,  $R^b$ ,  $R^c$  */
      end
       $F_a \leftarrow \frac{F_a}{2Z}$ 
       $F_b \leftarrow \frac{F_b}{2Z}$ 
      if  $Rand < (1 + e^{F_a - F_b})^{-1}$  then  $P_a \leftarrow P_b$ ;
      if  $t > 0.2Gens$  then
        /* keep track of the average number
        of cooperations */;
      end
    end
  end
end
  
```

**Extended Data Figure 6 | Pseudo code for the Monte Carlo simulations used to calculate the cooperation index under each norm.** Given the large number of norms considered, we used  $Runs = 15$

and  $Gens = 1.5 \times 10^4$  in Figs 2–4 and Extended Data Figs 1 and 2, and  $Runs = 50$  and  $Gens = 10^5$  in Extended Data Figs 4 and 5.



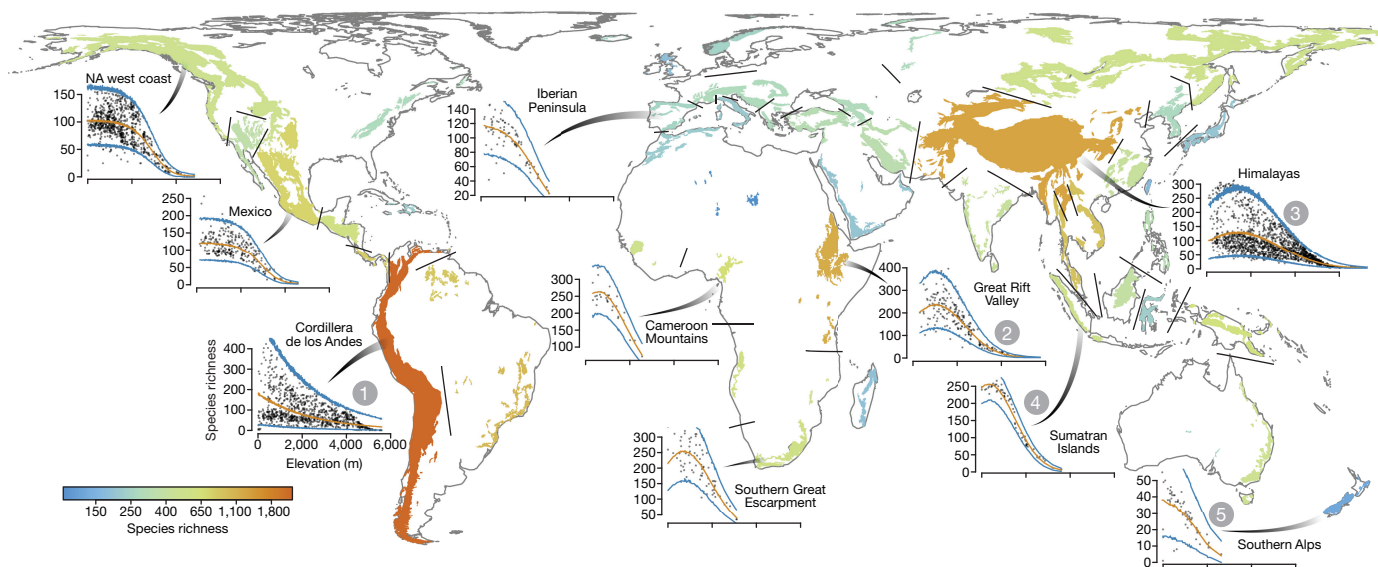
# Global elevational diversity and diversification of birds

Ignacio Quintero<sup>1</sup> & Walter Jetz<sup>1,2</sup>

Mountain ranges harbour exceptionally high biodiversity, which is now under threat from rapid environmental change. However, despite decades of effort, the limited availability of data and analytical tools has prevented a robust and truly global characterization of elevational biodiversity gradients and their evolutionary origins<sup>1,2</sup>. This has hampered a general understanding of the processes involved in the assembly and maintenance of montane communities<sup>2–4</sup>. Here we show that a worldwide mid-elevation peak in bird richness is driven by wide-ranging species and disappears when we use a subsampling procedure that ensures even species representation in space and facilitates evolutionary interpretation. Instead, richness corrected for range size declines linearly with increasing elevation. We find that the more depauperate assemblages at higher elevations are characterized by higher rates of diversification across all mountain regions, rejecting the idea that lower recent diversification rates are the general cause of less diverse biota. Across all elevations, assemblages on mountains with high rates of past temperature change exhibit more rapid diversification, highlighting the importance of climatic fluctuations in driving the evolutionary dynamics of mountain biodiversity. While different geomorphological and climatic attributes of mountain regions have been pivotal in determining the remarkable richness gradients

observed today, our results underscore the role of ongoing and often very recent diversification processes in maintaining the unique and highly adapted biodiversity of higher elevations.

Consensus about patterns and processes underlying the variation in species richness within and among mountain systems remains elusive<sup>2,5</sup>. On mountains, communities and abiotic conditions change rapidly over short distances, with greater elevational than lateral turnover in species composition<sup>6</sup>. This has complicated a synthetic understanding of montane diversity, because the high elevational turnover rates obfuscate straightforward accounting for the effects of scale and data non-independence. The rapid turnover in communities and conditions has also constrained the reconciliation of studies based on ‘alpha’ diversity (local richness) from single localities with those estimating ‘gamma’ richness (regional richness) summed for elevational bands of differing area<sup>2,7</sup>. Furthermore, estimates of elevational diversity are often complicated by the variation in sampling effort, sample size and human impact with respect to elevation<sup>7</sup>. For example, even in the comprehensively sampled mountains of Switzerland, field surveys consistently underestimate bird diversity towards higher elevations by an average of 29% (95% CI = 5.3–64%) above 1,850 m (Extended Data Fig. 1 and Supplementary Information). Here, we overcome these limitations by combining expert information on lateral distributions for all



**Figure 1 | Elevational gradients of avian species richness across mountain systems.** Variation in raw assemblage richness is shown for 10 out of the 46 mountain systems ( $n$  assemblages for each plot from left to right: 893, 301, 1,025, 80, 32, 73, 204, 69, 61, 1,816). The orange and purple lines correspond to the mean and 95% interval, respectively, of the posterior predictive negative binomial distribution regression.

All subplots are scaled so that they have the same elevation range. Five mountain systems are highlighted with numbers and followed through the subsequent figures. Black straight lines are visual guides to delineate the separate mountain systems. Continental coastline was downloaded from Natural Earth. NA, North America.

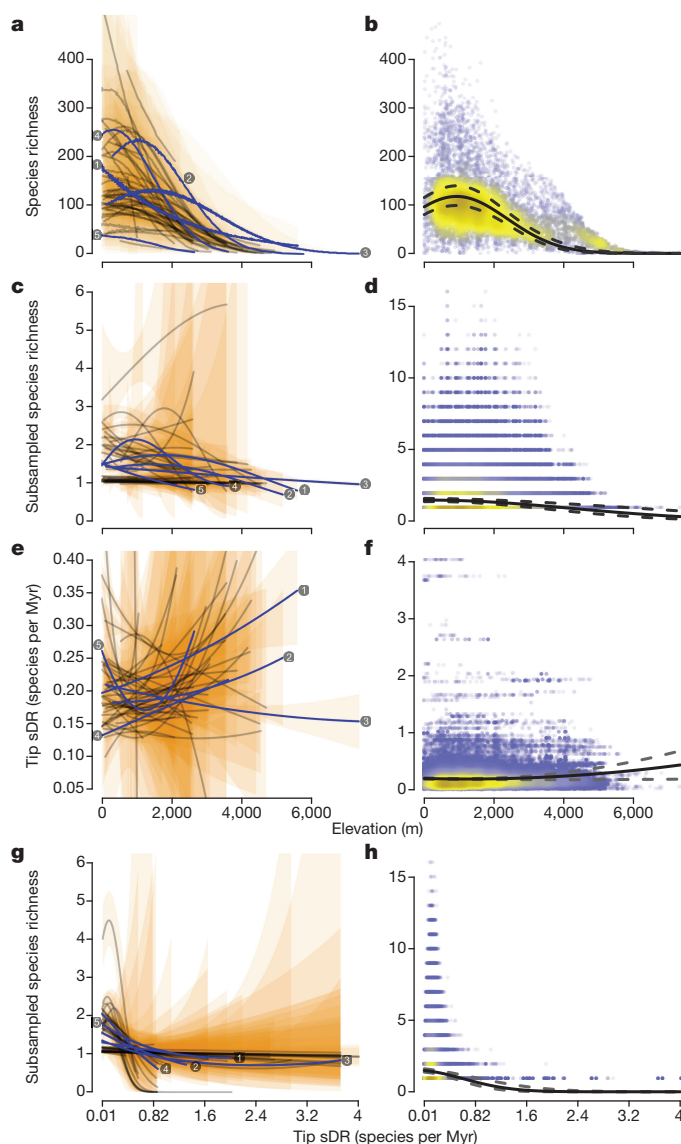
<sup>1</sup>Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, Connecticut 06520, USA. <sup>2</sup>Department of Life Sciences, Imperial College London, Silwood Park, Ascot, Berkshire SL5 7PY, UK.

birds with a new global compilation of mountain-range-specific data on elevational limits (see Methods and Supplementary Table 1) and use a novel statistical sampling procedure to characterize elevational richness gradients.

For the 46 major mountain systems of the world that represent distinct ‘evolutionary arenas’<sup>1</sup>, given their bird species composition<sup>8</sup> (Fig. 1, Extended Data Fig. 2a), we characterized bird assemblages in trapezoidal prisms of approximately 110 km lateral<sup>9</sup> and 500 m elevational extent, in accordance with the three-dimensional accuracy of underlying data (Extended Data Fig. 3; see Methods and Supplementary Information). This sampling resulted in 8,410 local assemblages containing 8,470 bird species (roughly 85% of all bird species) that capture the full elevational and lateral variation in diversity in each distinct mountain system (Fig. 1, Extended Data Fig. 2c). Compared to previously reported elevational transect information (see Supplementary Table 2), these data allow a globally consistent assemblage characterization that is robust to the typical sampling biases induced from variation in effort and sampling area along elevational bands (Extended Data Fig. 4; see Methods and Supplementary Information). We find considerable variation in patterns of elevational species richness among mountain systems, with a low elevation plateau as the most common pattern (Figs 1, 2a and Supplementary Figs 1–46). When we consider all assemblages of the world jointly, species richness falls within a triangular space along elevation (Fig. 2b), where maximum species richness decreases predictably towards higher elevations. Considering mountain systems as replicates in a Bayesian generalized hierarchical model suggests that the global elevational gradient in avian species richness is unimodal (Fig. 2b).

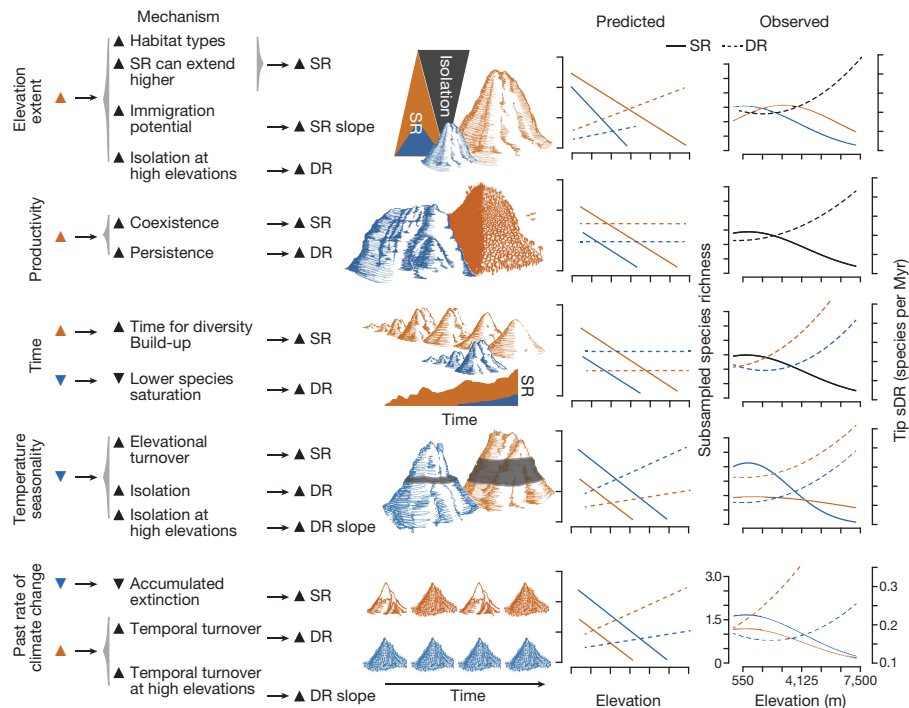
However, patterns of assemblage-based metrics, such as richness or summaries of species attributes, are most strongly influenced by wider-ranging species<sup>1,10</sup>. This issue impedes interpretation when species are not controlled for their differing contribution to aggregate metrics. Here, we address this unevenness using a random subsampling approach that provides three-dimensional range-size-controlled (or ‘subsampled’) estimates of assemblage metrics (see Methods). The elevational patterns of subsampled species richness (sSR; Fig. 2c) control for differing species range sizes (and their tendency to inflate a ‘mid-domain’ richness peak<sup>11</sup>) and enable inferences about differential diversification and/or dispersal of species in mountain systems that are not biased with regard to range size. We observe a much weaker and more linear elevational decrease in sSR than in species richness, with a low-elevation plateau followed by a decay (Fig. 2d). This pattern suggests that the frequently observed mid-elevation peak in raw species richness<sup>2,3</sup> derives from multiple occurrences of the same species at those elevations and is not particularly informative regarding the evolutionary processes that underlie this pattern.

The fundamental causes of species richness gradients are ultimately differences in speciation, extinction and migration<sup>12</sup>. The reduction in diversity at higher elevations may result from lower speciation or higher extinction rates (that is, diversification rates decrease with elevation). This may be due to smaller area (or younger age) affording fewer opportunities for speciation in space (or time) or increasing extinction risk<sup>13</sup>, or to lower temperatures and productivity depressing rates of speciation and population persistence, respectively<sup>4,14</sup>. Alternatively, diversification rates may be decoupled from the elevational richness gradient if, for instance, there are elevational differences in resource availability<sup>5</sup> and/or time of colonization<sup>12</sup>. Such decoupling could also arise when higher-elevation assemblages are mostly caused by immigration, as, for example, in the Himalayan avifauna<sup>15</sup>. Immigrants may in turn capitalize on new ecological opportunities by radiating rapidly, potentially enhanced by climate-change-induced perturbations that cause repeated extinctions followed by recent re-immigration and/or speciation, consistent with the idea of ‘ephemeral speciation’<sup>16</sup>. Thus, in an extreme scenario, higher elevations may contrast with more stable areas at lower elevations by having both higher recent diversification rates and lower species richness<sup>4</sup>.



**Figure 2 | Mountain-specific and worldwide patterns of elevational species richness gradients.** **a**, Posterior mean (95% posterior prediction intervals as orange shades) for raw species richness (SR) along elevation for the 46 mountain systems. For **c**, **e** and **g** the posterior mean (black/purple lines) and 95% credible intervals (CI; orange shades) are shown for each mountain. For **b**, **d**, **f** and **h**, global patterns using mountain systems as replicates in a hierarchical model are shown with solid and dashed lines representing the posterior mean and 95% CI, respectively. For **d**, **f** and **h**, localities are used as random effects. **b**, Best global model for species richness along elevation. **c**, Subsampled species richness (sSR) along elevation. **d**, Global monotonic decrease along elevation for sSR. **e**, Tip sDR along elevation. **f**, Global pattern of tip sDR along elevation. **g**, Effects of tip sDR on sSR. **h**, Global relationship between tip sDR and sSR. The focal mountains from Fig. 1 are highlighted in purple and labelled accordingly. The colour gradient in panels **b**, **d**, **f** and **h** represents the density of points, with warmer colours having more overlapping points and colder colours fewer. **a**, **b**,  $n = 8,410$  assemblages; **c**–**h**,  $n = 212,981$  subsampled assemblages.

Here, we integrate a global avian time-calibrated phylogenetic tree<sup>17</sup> to characterize elevational assemblages by their diversification rates using a species-level metric, tip diversification rate (tip DR, defined as the inverse equal splits rate<sup>17</sup>). When compared to related present-day tip rate estimates, tip DR provides higher resolution among recently diversified clades (Extended Data Fig. 5 and Supplementary Information). Unlike clade-level methods, and key for diversity gradient analyses, this metric allows the combination of the signal from



**Figure 3 | Hypotheses, predictions and results for explaining mountain level variation in species richness and diversification rates.** Diversification rate (DR) represents a general form of tip sDR. Mountain characteristics can affect the intercept or the slope of the relationship. Orange colour represents higher values for a given variable and purple colour represents lower values, as depicted in the cartoon drawings. Significant ‘Observed’ effects are shown as separate purple and orange

lines, whereas relationships that are non-significant are shown as a single black line. The specific positions of ‘Observed’ purple and orange lines depict the lower and upper 95% credible interval, respectively, for the effect of each covariate (see Supplementary Information). All ‘Observed’ subplots share the same axes ranges. Hierarchical models used  $n = 212,981$  subsampled assemblages. SR, species richness.

each component species to characterize the diversification rates of assemblages (Extended Data Fig. 6 and Supplementary Information). For example, assemblages composed of species resulting largely from recent radiations are characterized by an elevated tip DR.

As with richness, controlling for range size effects on assemblage tip DR through subsampling (tip sDR) ensures that all species are represented equally and provides an interpretation that is more relevant to evolutionary processes. Elevational patterns of tip sDR show strong variability among mountain systems (Fig. 2e and Supplementary Figs 1–46), but are globally characterized by a consistent (although weak) increase with elevation (Fig. 2f). However, tip sDR values are predominantly negatively related to sSR for all mountains (Fig. 2g and Supplementary Figs 1–46) and, globally, tip sDR shows a negative relationship with sSR (Fig. 2h). These findings lead us to reject the hypothesis that lower recent diversification rates drive the decrease in number of species towards higher elevations. Strong habitat differences among elevational zones<sup>18</sup> may often result in high-elevation immigrants sharing a diversification rate signal with close relatives from distant regions with similar climates, rather than from nearby lower elevations<sup>19</sup>. The isolation and heterogeneity of higher-elevation landscapes may facilitate rapid radiations of immigrant clades<sup>20</sup>, thereby increasing present-day estimates of diversification rates at higher elevations. Furthermore, although recent speciation events at high elevations increase diversification rates, they do not contribute greatly to the build-up of local diversity if they predominantly involve populations that have remained laterally allopatric within mountain regions.

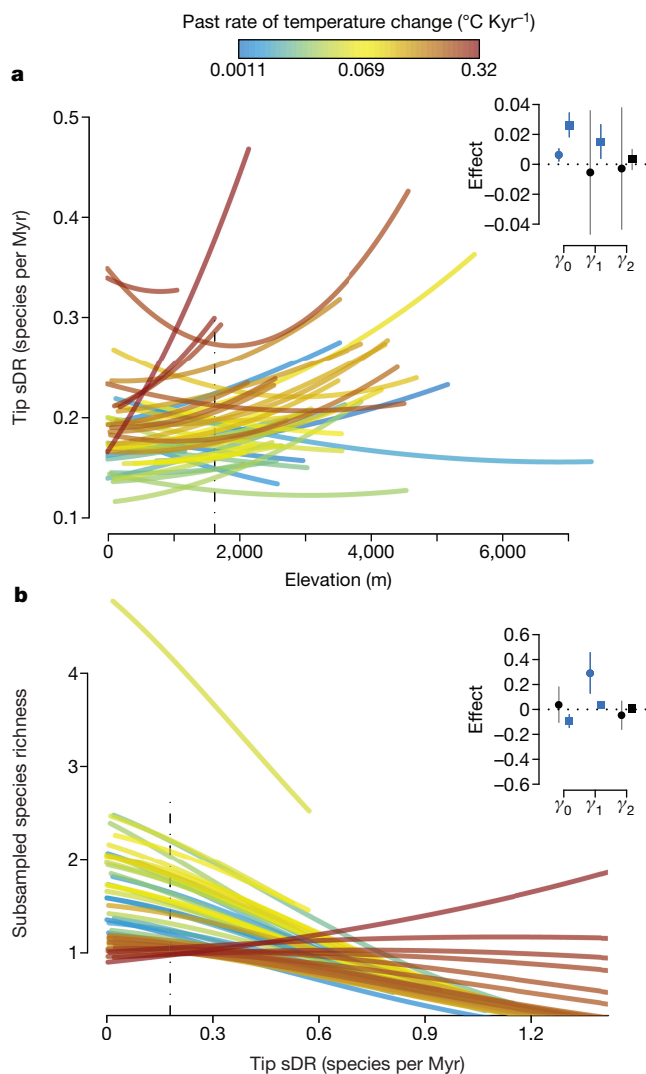
We explored how the elevational extent of mountain systems, their productivity, mountain age (time since uplift), seasonality and past rates of climate change affect the interplay between elevation and the species richness (sSR) and diversification rates (tip sDR) of assemblages (Fig. 3). Taller mountain ranges tend to be larger in area and harbour a greater variety of habitats, both of which result in greater opportunities for immigration, within-region isolation and *in situ* diversification<sup>1,21</sup>.

We find that, consistent with their greater areal extent and region-wide richness (Extended Data Fig. 7), taller mountain systems have more speciose assemblages and their unimodal richness peak is found at higher elevations (Fig. 3). Nonetheless, trends in elevational richness and in diversification rates are otherwise similar to those of lower-altitude mountains (Fig. 3, Extended Data Figs 8, 9). These results reaffirm the role of combined areal and elevational extent of a mountain region in shaping how species richness varies along elevation<sup>7</sup>.

Regional differences in available energy (productivity) may influence both the species richness and characteristic diversification rates of an assemblage, primarily by setting a ‘carrying capacity’ on the number of coexisting species<sup>5,22</sup> and by facilitating population persistence and thereby decreasing the probability of extinction of species<sup>23</sup>. However, regional productivity had no effect on either sSR or tip sDR along elevation. We also investigated whether mountain age (Supplementary Table 3) had an effect on present-day richness following the time-for-speciation effect hypothesis<sup>24</sup>. The uplift of a mountain system provides new habitats for neighbouring clades to colonise. With longer stretches of time, clades have more opportunities to diversify *in situ*, thereby increasing regional species richness<sup>25</sup>. While we detect no effect of mountain age on species richness, we find that assemblages on older mountain systems have higher diversification rates, especially towards higher elevations (Fig. 3, Extended Data Fig. 9). We speculate that older mountains could have acted as long-standing refuges for radiating lineages during extreme environmental perturbations.

Whereas species that experience large temperature changes at higher latitudes tend to be physiological and ecological generalists with wider elevational ranges, those in more aseasonal, tropical mountain systems have narrower elevational ranges<sup>26</sup>. This narrower elevational specialization in low-seasonality areas should facilitate isolation and associated opportunities for speciation, with both leading to higher assemblage richness, particularly at lower-to-mid elevations, compared to highly seasonal mountains<sup>27</sup>. We find that mountain systems with greater





**Figure 4 | Effect of past rates of climate change on the interplay between diversification rates and species richness along elevation.** The results are based on global multilevel models with mountain system as a random effect and mountain-level covariates. **a**, Mountain systems with higher rates of past climate change have higher tip sDR, and increasingly so towards higher elevations. **b**, They also exhibit a more positive relationship between tip sDR and sSR, while the relationship is strongly negative in climatically stable mountain regions. Lines correspond to mountain systems. The inset plots show the effect of the covariate on the intercept ( $\gamma_0$ ), the linear ( $\gamma_1$ ) and the quadratic coefficient ( $\gamma_2$ ; posterior average and 95% CI); circles correspond to SR and tip DR, squares to sSR and tip sDR; blue coloured effects correspond to coefficients where the 95% CI does not overlap with 0. The vertical dashed black line corresponds to the average of the x-axis (see Supplementary Information and Extended Data Figs 8, 9.) Hierarchical models based on 42,526 subsampled assemblages across 46 mountain systems.

seasonality have much lower assemblage richness at low elevations than their more aseasonal counterparts (Fig. 3), consistent with the well-known latitudinal gradient in species richness<sup>1,25</sup> (Extended Data Figs 8, 9). Notably, this effect is reversed at the highest elevations, where more seasonal mountain systems maintain relatively higher level of richness (Fig. 3). In contrast to our predictions, more seasonal mountains are characterized by higher diversification rates (Fig. 3).

Regions that—like higher elevations—have experienced large climatic fluctuations are expected to undergo higher temporal turnover of species<sup>28</sup>. For example, glaciation cycles have caused differential extinctions along the latitudinal gradient, spurring recent re-colonizations and radiations<sup>28,29</sup>. Consequently, we predict that

more climatically dynamic regions will show lower richness across all elevations while having an overall elevated tip sDR signal. By contrast, more stable areas should have lower temporal species turnover, thus retaining lineages from ancient radiations that exhibit lower average rates of diversification. In concordance with both sets of our expectations, richness along elevation is lower and absolute levels and elevational increases in diversification rates are much greater in mountains that have experienced higher rates of past climate change (Figs 3, 4a). Notably, in more stable regions, richness tends to be negatively associated with diversification rates, with speciose assemblages having the lowest levels of tip sDR (Fig. 4b). By contrast, in climatically more variable mountain systems, species richness and diversification rates range from weakly to slightly positively related. Past rates of climate change emerge as the most consistently supported hypothesized determinant of elevational variation in diversification rates among mountain systems (Fig. 3 and Supplementary Information), asserting the significance of climatic fluctuations for geographic gradients of speciation and diversification.

In summary, we consistently find that recent diversification rates are highest where current species richness is lowest, and that this effect is stronger in more climatically stable regions. These findings suggest that at the spatial grain analysed here, the signal of diversification rates in generating biodiversity is effaced over geological timescales. These results confirm previous findings<sup>17,30–32</sup> that contrast with hypotheses that attempt to explain modern global patterns on the basis of origination rates, such as the latitudinal diversity gradient. Instead, low extinction rates because of long-term environmental stability seem to play a bigger role in the maintenance of regional species richness<sup>25</sup>. The inadequate capture of extinction rates from present-day-only taxa limit a more thorough understanding of past diversification dynamics, but our findings highlight the importance of a more historically informed understanding of diversity gradients. Mountain ranges harbour a large fraction of extant biodiversity that is often uniquely adapted and is now under immense threat from climate change, calling for more detailed and taxonomically comprehensive monitoring and characterization of montane species. Our findings illustrate the distinct evolutionary implications brought about by the specific geomorphology, isolation, and climatic fluctuations of each mountain region, and highlight the need to safeguard both the past outcome and future of their diversification dynamics.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 January 2017; accepted 31 January 2018.

Published online 21 February 2018.

1. Jetz, W. & Fine, P. V. A. Global gradients in vertebrate diversity predicted by historical area-productivity dynamics and contemporary environment. *PLoS Biol.* **10**, e1001292 (2012).
2. McCain, C. M. Global analysis of bird elevational diversity. *Glob. Ecol. Biogeogr.* **18**, 346–360 (2009).
3. Rahbek, C. The relationship among area, elevation, and regional species richness in neotropical birds. *Am. Nat.* **149**, 875–902 (1997).
4. Graham, C. H. *et al.* The origin and maintenance of montane diversity: integrating evolutionary and ecological processes. *Ecography* **37**, 711–719 (2014).
5. Price, T. D. *et al.* Niche filling slows the diversification of Himalayan songbirds. *Nature* **509**, 222–225 (2014).
6. Terborgh, J. Distribution on environmental gradients: theory and a preliminary interpretation of distributional patterns in the avifauna of the Cordillera Vilcabamba, Peru. *Ecology* **52**, 23–40 (1971).
7. Nogués-Bravo, D., Araújo, M. B., Romdal, T. & Rahbek, C. Scale effects and human impact on the elevational species richness gradients. *Nature* **453**, 216–219 (2008).
8. Körner, C. *et al.* A global inventory of mountains for bio-geographical applications. *Alp. Bot.* **127**, 1–15 (2016).
9. Hurlbert, A. H. & Jetz, W. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl Acad. Sci. USA* **104**, 13384–13389 (2007).
10. Jetz, W. & Rahbek, C. Geographic range size and determinants of avian species richness. *Science* **297**, 1548–1551 (2002).



11. Colwell, R. K. & Lees, D. C. The mid-domain effect: geometric constraints on the geography of species richness. *Trends Ecol. Evol.* **15**, 70–76 (2000).
12. Kozak, K. H. & Wiens, J. J. Niche conservatism drives elevational diversity patterns in Appalachian salamanders. *Am. Nat.* **176**, 40–54 (2010).
13. Kisel, Y., McInnes, L., Toomey, N. H. & Orme, C. D. L. How diversification rates and diversity limits combine to create large-scale species-area relationships. *Phil. Trans. R. Soc. Lond. B* **366**, 2514–2525 (2011).
14. Hawkins, B. A. *et al.* Energy, water, and broad-scale geographic patterns of species richness. *Ecology* **84**, 3105–3117 (2003).
15. Päckert, M. *et al.* Horizontal and elevational phylogeographic patterns of Himalayan and Southeast Asian forest passerines (Aves: Passeriformes). *J. Biogeogr.* **39**, 556–573 (2012).
16. Rosenblum, E. B. *et al.* Goldilocks meets Santa Rosalia: an ephemeral speciation model explains patterns of diversification across time scales. *Evol. Biol.* **39**, 255–261 (2012).
17. Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Mooers, A. O. The global diversity of birds in space and time. *Nature* **491**, 444–448 (2012).
18. Körner, C., Paulsen, J. & Spehn, E. M. A definition of mountains and their bioclimatic belts for global comparisons of biodiversity data. *Alp. Bot.* **121**, 73–78 (2011).
19. Barker, F. K., Burns, K. J., Klicka, J., Lanyon, S. M. & Lovette, I. J. New insights into New World biogeography: an integrated view from the phylogeny of blackbirds, cardinals, sparrows, tanagers, warblers, and allies. *Auk* **132**, 333–348 (2015).
20. Smith, B. T. *et al.* The drivers of tropical speciation. *Nature* **515**, 406–409 (2014).
21. Kisel, Y. & Barraclough, T. G. Speciation has a spatial scale that depends on levels of gene flow. *Am. Nat.* **175**, 316–334 (2010).
22. Marshall, C. R. & Quental, T. B. The uncertain role of diversity dependence in species diversification and the need to incorporate time-varying carrying capacities. *Phil. Trans. R. Soc. Lond. B* **371**, 20150217 (2016).
23. Wright, D. H., Currie, D. J. & Maurer, B. A. in *Species Diversity in Ecological Communities: Historical and Geographical Perspectives* (Univ. Chicago Press, 1993).
24. Stephens, P. R. & Wiens, J. J. Explaining species richness from continents to communities: the time-for-speciation effect in emydid turtles. *Am. Nat.* **161**, 112–128 (2003).
25. Fine, P. V. A. Ecological and evolutionary drivers of geographic variation in species diversity. *Annu. Rev. Ecol. Evol. Syst.* **46**, 369–392 (2015).
26. Janzen, D. H. Why mountain passes are higher in the tropics. *Am. Nat.* **101**, 233–249 (1967).
27. Cadena, C. D. *et al.* Latitude, elevational climatic zonation and speciation in New World vertebrates. *Proc. R. Soc. Lond. B* **279**, 194–201 (2012).
28. Hewitt, G. The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913 (2000).
29. Wallis, G. P., Waters, J. M., Upton, P. & Craw, D. Transverse alpine speciation driven by glaciation. *Trends Ecol. Evol.* **31**, 916–926 (2016).
30. Weir, J. T. & Schluter, D. The latitudinal gradient in recent speciation and extinction rates of birds and mammals. *Science* **315**, 1574–1576 (2007).
31. Weir, J. T., Bermingham, E. & Schluter, D. The Great American Biotic Interchange in birds. *Proc. Natl Acad. Sci. USA* **106**, 21737–21742 (2009).
32. Smith, B. T., Seeholzer, G. F., Harvey, M. G., Cuervo, A. M. & Brumfield, R. T. A latitudinal phylogeographic diversity gradient in birds. *PLoS Biol.* **15**, e2001073 (2017).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank R. E. Ricklefs, P. V. A. Fine, T. Price, C. D. Cadena, J. Beck, E. Spriggs, N. Upham and R. Freckleton for manuscript comments; members of the Future Earth Global Mountain Biodiversity Assessment, including C. Körner, E. Spehn, M. Fischer, and D. Payne for feedback; B. Klempay, A. Houston and A. Ranipeta for help collecting the elevational data; and the ‘Monitoring Häufige Brutvögel (MBH)’ project of the Vogelwarte Sempach (M. Kery) for sharing data on the Swiss breeding bird survey from 2007. We acknowledge the World Climate Research Programme’s Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups for producing and making available their model output. This material is based upon work supported by NSF grants DGE-1122492, DEB-1441737, and DBI-1262600.

**Author Contributions** I.Q. and W.J. designed the research and wrote the manuscript. I.Q. conducted the analyses.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to I.Q. ([ignacio.quintero@yale.edu](mailto:ignacio.quintero@yale.edu)).

**Reviewer Information** *Nature* thanks A. Antonelli and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**Elevation biases in richness estimates: Swiss bird survey.** To demonstrate the level of bias that may arise along elevational gradients from local surveys of assemblage richness estimates, we used one of the most comprehensive biodiversity datasets available, the Swiss breeding bird survey ('Monitoring Häufige Brutvögel') conducted in 2007 by the Vogelwarte Institute (<http://www.vogelwarte.ch/en/projects/monitoring/monitoring-common-breeding-birds.html>). This systematic inventory surveyed all breeding birds during three summer expeditions across 267 sampling sites of 1 km<sup>2</sup> throughout the country. The repeated and standardized sampling design allows the use of occupancy models, which enable quantification of detectability and potential richness biases along elevation. We applied *N*-mixture occupancy models that model detection bias directly from the multi-species occupancy model with data augmentation as previously described<sup>33</sup>, which treats the observed species as a subsample of a larger, unobserved, community of species<sup>34</sup>. We modelled the effect of elevation on species richness, taking into account observation bias and measurement error and propagating the uncertainty within a Bayesian framework using the *rjags*<sup>35</sup> package for R<sup>36</sup>. This model estimates a posterior probability distribution for the number of species at each sampling site and the effect of elevation on species richness and detection probability.

**Delimiting mountain systems.** The major mountain systems tend to represent geographically, biotically and environmentally distinct entities. In many cases their different geological and biogeographic histories and often highly different biota qualify them as separate 'evolutionary arenas'<sup>1</sup>, that is, as biologically significant replicates for analysis. A recent effort by the Global Mountain Biodiversity Assessment (GMBA, <http://www.mountainbiodiversity.org>) provided an inventory of the World's mountain regions based on expert delineation and terrain ruggedness<sup>8</sup>. This inventory identified 46 broad-scale mountain regions across the five continents (Extended Data Fig. 2a). To evaluate the biological independence of this delineation, we used species distribution data for birds (see below) to develop species lists for all mountain ranges and quantified their similarity in species composition based on the Jaccard index of similarity (species shared/all species present)<sup>37</sup>. We found that overall only a small proportion of species is shared among the 46 regions (median similarity of 0.47% ± 10.37% (s.d.)).

To assess whether a mountain region delineation designed to quantitatively maximize biotic region difference offered a strong improvement, we developed a second delineation as follows. First, we estimated for each ~110 km × 110 km terrestrial grid cell the average, minimum, maximum and range of elevation using the EarthEnv-DEM90, a 90 m Digital Elevation Model<sup>38</sup>. We used affinity propagation clustering (APC)<sup>39</sup> in the *apcluster* package<sup>40</sup> for R v.3.1.3<sup>36</sup> to create a similarity distance matrix between grid cells and divide them into 'highlands' and 'lowlands'. Subsequently, we applied the same clustering algorithm to cells identified as 'highlands' to group into mountain ranges following geographic distances and continuity (that is, cells identified as 'highlands' not separated by cells identified as 'lowlands'). We then used avian distribution data (see below) to group the result from the previous analysis according to their similarity in species composition based on the Jaccard index. This similarity matrix was then used as input for clustering into biologically independent mountain systems using APC. The resulting number of mountain regions depends on the quantile threshold value specified. The default value, which corresponds to the median of the dissimilarity matrix, gave too coarse mountain regions when visually inspected, so we used a lower cut-off value. This approach yielded 50 different regions that showed strong overlap with the GMBA delineation with a median species similarity among them that was only slightly higher (1.63% ± 13.29% (s.d.)). For overall generality and comparability with subsequent work, we thus retained the GMBA major mountain regions as our units of analysis.

**Distributions and elevational ranges.** We used data on breeding distributions compiled from the best available sources for a given broad geographical region or taxonomic group<sup>17,41</sup> totalling 9,993 species (for individual maps, see <https://mol.org>). These ranges were previously validated to have minimum (<5–10%) false presences at spatial grains larger than approximately 100 km<sup>9</sup>. We compiled a database of bird elevational ranges based on a total of 318 published sources and consisting of 27,840 species/mountain-range-specific entries. For full data and source-specific information, see Supplementary Table 1, also available at <https://mol.org/downloads>. When information was available, we used separate elevational ranges for each mountain system, incorporating different elevational ranges for widely distributed species. When different elevational ranges of a species were available for the same mountain range, we used the minimum and maximum amongst all given ranges. We followed a standardized rule set to threshold frequent adjectives used for characterizing elevational ranges. We did not include within a species' elevational range elevations for which adjectives such as 'rarely' or 'infrequently' were used, assuming these records are possibly vagrants and do not represent the environmental habitat of the species. By contrast, we did include elevations for which adjectives such as 'common', 'frequently' and 'often' were used

(see Supplementary Table 1). Species labelled as 'Coastal/Marine' were specified a maximum elevational range of 150 m. For the Cordillera de los Andes we used country-specific elevational ranges when available, using first *The Handbook of the Birds of the World*<sup>42</sup> but mostly published field guides for Andean countries (complete information on sources is available in Supplementary Table 1). We note that, usually, the information in *The Handbook of the Birds of the World* reflected information from national or regional field guides. We fitted local polynomial regressions (Loess curves<sup>43</sup>) for the minimum and maximum elevation across latitude for each Andean species, allowing us to interpolate the elevational range for every species at any latitude. We used the latitudinal midpoint of each country in the regression except if the country was either the northern or southern extreme of the species distribution, in which case we used the respective northern or southern latitude limit of the species geographic range.

For the 331 species for which we did not find information about their elevational ranges, we assessed their elevational range based on their expert geographic distribution and carefully vetted sampled localities using the species-refinement tool in Map of Life (<https://mol.org>)<sup>44</sup>. Finally, we quality refined this information further using the 110-km gridded expert breeding range distributions (see above) and maximum and minimum elevation from EarthEnv-DEM90, a 90-m resolution Digital Elevation Model (DEM)<sup>38</sup>. For every species, we used the maximum elevational range possible across its distribution to constrain its elevational range. Handling different elevational sources compelled us to contemplate a reliable elevational 'grain', where we minimize false absences and false presences. Relatively limited, but certainly existing, data incongruences among sources suggested a 500-m elevation grain as a good compromise between sufficient detail and a minimization of false absences as addressed in the validation section below. Sensitivity analyses into the grain choice show that assuming a less conservative elevational grain of 300 m has negligible effects on elevational patterns of richness and tip DR and does not significantly affect parameter estimates and results (Extended Data Fig. 10 and Supplementary Information).

**Elevation detection bias and assemblage richness comparison and validation.** We conducted an analysis to document how the heterogeneity of field-based richness data complicates a straightforward synthesis of elevational richness gradients and how such data compare to the standardized estimates derived by our approach of elevationally refined range maps. For this, we compiled 37 local surveys of species richness along elevation gradients of the sort used in previous richness gradient comparisons, comprising a total of 370 elevation locations with richness data and associated estimates of survey effort, estimated as the approximate time spent on each elevation location (Supplementary Table 2). We then compared these data with the richness estimates obtained through our approach (see above) and related their differences to sampling effort.

To test the influence of sampling effort when comparing richness estimates from field surveys and those derived from our sampling units, we conducted a hierarchical mixed effects linear regression. We used estimated species richness as a linear predictor of the richness derived from field surveys, using each separate field survey as a random effect. If sampling effort mostly drives the observed differences between the richness estimates, we expect the slopes to increase with increasing sampling effort, reaching values close to 1 for highly sampled surveys. Thus, we added the logarithm of sampling effort as a predictor of the slopes for each transect regression. We opted to use the logarithm of sampling effort because richness accumulation curves with increasing effort are known to be nonlinear: richness estimates increase rapidly at first but then slow down, eventually reaching an asymptote<sup>45</sup>. We ran this regression in a Bayesian framework using Integrated Nested Laplace Approximation (INLA) using the R-INLA package<sup>46</sup> for R<sup>36</sup>.

**Biodiversity data sampling.** We divided the world into three-dimensional trapezoids of 110-km lateral and 500-m elevational extent, reflecting the assumed accuracy of our species range data along three dimensions (Extended Data Fig. 3). These units are more densely spaced along a steep slope than a plateau, and divide a mountain range into separate comparable units. This sampling allows us to combine our two sources of information while maintaining appropriate resolution across three-dimensional space: first, expert distribution maps (that is, polygons) give us information along longitude and latitude, and second, species elevational limits allow us to determine its position along elevation. Species region-specific elevational information enables denser sampling along mountain slopes, because it allows us to differentiate between, for instance, lowland and highland assemblages that would otherwise be lumped together. We then randomly selected one point within the area of each trapezoid and intersected it with all qualifying bird presences to derive an alpha diversity estimate for that unit. The sampling was implemented using the *raster*<sup>47</sup> and *rgdal*<sup>48</sup> packages in R<sup>36</sup>. The sampling returned a total of 21,655 assemblages across the world (Extended Data Fig. 2b) and 8,410 assemblages within our delimited mountain systems (Extended Data Fig. 2c) with a median of 71 per mountain region (minimum 3 in the Central Australian mountains, maximum 1,816 in the Himalayas).

**Species richness along elevation.** To explore how species richness varies with elevation we used three different models to address mean richness ( $\mu$ ) along elevation: following a linear relationship,

$$\log(\mu_i) = \beta_0 + \beta_1 x_i \quad (1)$$

a low-elevation plateau (that is, sigmoidal) relationship,

$$\log(\mu_i) = \beta_0 / (1 + e^{(\beta_1 + \beta_2 x_i)}) \quad (2)$$

or a mid-elevation peak (that is, unimodal) relationship,

$$\log(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \quad (3)$$

We first examined how species richness varies with elevation across each mountain system. For each mountain range, we fitted each of the models (equations (1)–(3)) in which species richness follows a negative binomial (NB) distribution using generalized linear models (GLMs):

$$\text{Species richness}_i \sim \text{NB}(\mu_i, k)$$

within a Bayesian framework using JAGS v.3.4.0 through the rjags package for R v.3.1.3<sup>35,36,49</sup>. We note that our definition of linear, low-elevation plateau and mid-elevation peak differs from that used in previous literature<sup>2</sup>. Our definition is based on statistically selecting the best fitting model, and is not defined by absolute elevation values. We estimated the Watanabe–Akaike information criterion (WAIC), which is preferable to similar alternatives because it averages over the posterior distribution rather than using only point estimates<sup>50</sup>, and selected the best model accordingly for each mountain system. To estimate a global pattern, we used the same three models for the mean (equations (1)–(3)) in a two-level generalized hierarchical model, in which the parameters for each mountain system come from a multivariate normal distribution:

$$\beta_m \sim \text{MVN}(\mu, \Sigma) \quad (4)$$

$$\Sigma = \begin{pmatrix} \sigma_0^2 & \dots & \rho\sigma_0\sigma_p \\ \vdots & \ddots & \vdots \\ \rho\sigma_0\sigma_p & \dots & \sigma_p^2 \end{pmatrix},$$

in which there are  $p$  hyper-parameter means such that  $\mu = \{\mu_0, \dots, \mu_p\}$ , with  $m$  mountain-system specific realizations  $\beta_m = \{\beta_{0m}, \dots, \beta_{pm}\}$ . We fitted several models in a Bayesian framework using INLA<sup>51</sup> through the R-INLA package for R<sup>46</sup>, starting with the most general model in which all mountain-specific parameters come from a multivariate normal distribution and are correlated between them, and following with simplifications of such (that is, of  $\Sigma$ ). We selected the best global model according to the WAIC.

**Bird diversification rates.** *Phylogenetic information and tip DR.* Phylogenetic information was obtained from ref. 17 by randomly sampling 1,000 trees from the posterior distribution of trees produced with the Hackett backbone, allowing us to capture phylogenetic uncertainty in subsequent analyses<sup>17,52</sup>. To measure diversification rates at each location, we used a species-level diversification rate<sup>17</sup> based on the equal-splits metric (tip DR). We calculated tip DR for each species across the posterior distribution of trees and estimated the harmonic average and standard deviation. Finally, for each assemblage we estimated diversification rate as the harmonic mean of the tip DR of the constituent species.

Furthermore, we compared tip DR to both traditional metrics addressing whole clades using a constant birth–death process and to BAMM<sup>53</sup>-derived tip rates of diversification and underline the advantages of using tip DR in spatial analyses of diversification (Extended Data Figs 5, 6 and Supplementary Information). *Diversification rates along elevation.* We explored how diversification rates vary along elevation for each mountain system using the same three models for the mean used for characterizing species richness along elevation (equations (1)–(3)), but allowing errors to follow a Gaussian distribution:

$$\text{Diversification rates}_i \sim N(\mu_p, \sigma^2)$$

Similarly, we estimated a global cross-mountain pattern of diversification rates along elevation using hierarchical models as with species richness. We then explored to what extent species richness patterns are coupled with diversification rates across mountain ranges. To this end, we estimated the effect of assemblage diversification rates on species richness for each mountain system using the multilevel modelling framework outlined above (equations (1)–(3)) with negative

binomial regressions. Finally, we constructed a global model using the hierarchical approach outlined in equation (4).

**Controlling for overrepresentation.** Up to this point we have described the use of aggregate measures such as local assemblage richness and diversification rates based on lists including all species expected to occur in an assemblage. Species lists alone are, however, limited in allowing macroevolutionary inference, as wide-ranging species drive the variation in aggregate metrics, such as assemblage richness or average species attribute values, much more than narrow-ranged species<sup>1,10</sup>. Inference that values all species equally, such as typically expected for macroevolutionary analyses or the case for phylogenetic analyses of diversification rates, requires accounting for this uneven contribution from the different species. For instance, estimating the effect of diversification rates that characterize assemblages on species richness could be systematically decoupled if in the former each species contributes equally (each species is counted once, independent of their distribution range) while wide-ranged species are over-represented in the latter. Here, we address this unevenness using a random subsampling approach that provides range-size controlled estimates of assemblage species richness and diversification rates to which each species contributes statistically equally.

**Random subsampling.** First, we extended our assemblage sampling to the whole world, that is, terrestrial regions outside our mountain systems, using all the 21,655 sampling locations shown in Extended Data Fig. 2b. Given this set of locations, we compiled a list of species present at each location. Let  $L = \{l_1, \dots, l_q\}$  be the set of  $q$  locations in space (that is,  $q = 21,655$  for this study) and let  $S_j$  be the set of locations in which species  $j$  is present such that  $S_j \subseteq L$ . Define  $r_j$  as the range size of species  $j$  such that  $r_j = |S_j|$ . Because species ranges can differ greatly in size and shape, wide-ranging species would be represented in more locations than narrow-ranging species. To control for this unequal representation, we first select the species with the smallest range (that is, the fewest presences across the set of locations), such that  $r_{\min} = \min\{r_1, \dots, r_{\text{total species}}\}$ .  $r_{\min}$  cannot be less than 1. It is easily seen that species who share the same minimum range size will have the smallest statistical effect on any estimates being inferred. Thus, for every species  $j$ , we sample uniformly and without replacement from the set  $S_j$  the same number of locations as the species with the minimum range. The first draw,  $y_{j1}$ , for species  $j$  has probability  $\Pr(y_{j1}) = \frac{1}{r_j}$ , the second draw, if applicable (that is,  $r_{\min} > 1$ ), has probability  $\Pr(y_{j2}) = \frac{1}{r_j - 1}$ , and so forth. We then define the non-empty set  $U_j = \{y_{j1}, \dots\}$  such that  $U_j \subseteq S_j$  and  $|U_j| = r_{\min}$ . Clearly, only using  $U_j$  for analyses makes each species contribute equally, but one would not be anywhere close to properly characterizing the spatial patterns. Thus, let  $U_{j1}$  be the first set of draws obtained above and repeat the procedure of randomly drawing  $r_{\min}$  locations from each  $S_j$ , obtaining  $U_{j2}$ . Finally, we repeat this random subsampling for  $i$  iterations to obtain the set of sets  $\{U_{j1}, \dots, U_{ji}\}$ .

For ease of understanding, let us now focus on a given location  $l_i$  at which a total of  $T_h$  species has been recorded. When we perform one iteration of the subsampling described above, the new set of species present,  $O_h$ , will be a subset of the total number of species, including the empty set  $\emptyset$  (that is,  $O_h \subseteq T_h$  and  $|O_h| \leq |T_h|$ ). Thus, given  $i$  iterations of random subsampling, we observe different subsets of species for a given locality,  $\{O_{h1}, \dots, O_{hi}\}$ . These subsets are not independent of each other since they come from the same larger set  $T_h$ . In spatial analyses, we are usually interested in some characteristic of the assemblage  $T_h$  in relation to other properties of  $T_h$  or of the location  $l_i$ . Let  $c_{T_h}$  be a property of  $T_h$ , say, species richness, such that  $c_{T_h} = |T_h|$ . We can then quantify a new measure, subsampled species richness or sSR, on the results of our random subsampling  $O_h$ . This yields a new set  $C_{O_h} = \{|O_{h1}|, \dots, |O_{hi}|\}$  comprised of the subsampled species richness for each iteration. For clarity, each location  $l_i$  has a distribution,  $C_{O_h}$ , of  $i$  values for subsampled species richness. This distribution has the desired advantage of being comprised of values where each species contributes statistically equally. However, the issue of non-independence between  $C_{O_h}$  values remains. This can be accommodated within a hierarchical model (sometimes called a mixed-effects model), in which the location  $l_i$  is used as a random effect. When used within a proper inference framework, we can propagate the inherent uncertainty of these distributions. This approach applies to any type of assemblage measure. For instance, if we want to quantify average tip DR across space, we take the harmonic average over the tip DR of the species in  $O_h$ , resulting in a distribution of subsampled average tip DR (tip sDR) for location  $l_i$ . Similarly, we use a hierarchical model in which  $l_i$  is used as a random effect.

We used simulations to show that there was no systematic bias in this subsampling approach (Supplementary Information).

We used 100 iterations, resulting in a frequency distribution of sSR and tip sDR for each locality. This process represents each species for the same number of localities, making their contribution to the estimates of diversification rates and species richness the same, thus avoiding pseudoreplication.



**Subsampled multilevel regression.** To explore how sSR and tip sDR vary along elevation, we take into account the full distribution at each locality using a two-level model for each mountain range. We use the models in equations (1)–(3) to model the mean change along elevation, in which the mean parameters for each locality  $h$  come from a multivariate normal distribution:

$$\beta_h \sim \text{MVN}(\mu, \Sigma).$$

We also modelled the effect of tip sDR on sSR using the same modelling framework. Again, we used a negative binomial distribution to model sSR as a response and a Gaussian distribution for tip sDR. As general and global models for each of the three relationships, we used equations (1)–(3) and we incorporated each mountain range  $m$  as an evolutionary replicate using a three-level hierarchical model:

$$\beta_{hm} \sim \text{MVN}(\mu_m, \Sigma_m)$$

$$\mu_m \sim \text{MVN}(\mu, \Sigma) \quad (5)$$

in which  $h = 1, 2, 3, \dots, n_m$  represents the  $n_m$  sampling units on each mountain range  $m$ , and  $\mu$  is the vector of the hyper means. We selected the optimal model as determined by the WAIC. These models allow us to fully account for variation in the data across all mountain systems when determining the global patterns.

**Explaining the variation in elevational richness gradients.** We explored several hypotheses regarding environmental and historical predictors of variation in elevational richness gradients among mountain regions (Supplementary Information). *Mountain characteristics.* We extracted area and elevational extent for every mountain system using the EarthEnv-DEM90 product<sup>38</sup>. We extracted mean annual net primary productivity (NPP) for each mountain system, averaged over 2000 to 2012, derived from MODIS17<sup>54</sup>. We estimated average annual temperature and average annual seasonality for each mountain system from the CRU CL 2.0 dataset, which consists of monthly means averaged over 1961–1990<sup>55</sup>. Because we expect higher mountains to have lower average temperatures, all other things being equal, we required a temperature measure that is independent of the elevational extent of the mountain system. Thus, we estimated the average temperature across a band surrounding each mountain system of ~30 km in width. We extracted absolute latitude for each mountain system using the weighted polygon centroid.

We collated geological information on the most likely event of orogeny for each mountain system (see Supplementary Table 3 for specific sources). Unfortunately, information on mountain uplift is not available for each of our mountain systems, and in some cases, there were different uplift pulses or different regions within a mountain system uplifted at different times. In these cases, we used the oldest and nearest known significant uplift pulses (see Supplementary Table 3 for details). Since some mountains are older than the estimated dates of the emergence of modern birds, we used the MRCA of the bird-tree, according to the latest phylogenetic analysis (that is, ~72.9 Mya) as the maximum time for a mountain system<sup>56</sup>.

Finally, to estimate the severity of past climatic fluctuations among mountain systems, we used spatial layers for past annual mean temperatures based on reconstructions provided by global climate models (GCMs). We used climatic reconstructions of the last interglacial (LIG; ~120–140 Kya), the last glacial maximum (LGM; ~22 Kya), and the mid-Holocene (MH; ~6 Kya). For the LIG we used the reconstruction from ref. 57. For the LGM and MH we used the Coupled Model Intercomparison Project 5 (CMIP5) multimodel ensemble from the World Climate Research Programme's Working Group on Coupled Modelling<sup>58</sup>. Specifically, for the LGM we averaged over the following GCMs: CCSM4, MIROC-ESM, and MPI-ESM-P. For the MH, we averaged over the following GCMs: BCC-CSM1-1, CCSM4, CNRM-CM5, HadGEM2-CG, HadGEM2-ES, IPSL-CM5A-LR, MIROC-ESM, MPI-ESM-P, and MRI-CGCM3. Supplementary Table 4 lists the specific model groups and GCMs. We downloaded the information directly from WorldClim<sup>59</sup> at a 10' resolution, or, in the case of LIG, at a 30' resolution and then aggregated using the average into a 10' resolution.

We estimated the rate of mean annual temperature change by taking the absolute value of the subtraction between LIG and LGM and between LGM and MH, and then dividing by the appropriate time interval for each grid cell. Subsequently, we used the harmonic average of these two rates for each grid cell. Finally, for each mountain system, we used the harmonic average across all intersecting grid cells as a region-wide characterization of past climate change.

While we collated the best available information for climatic and geological information, we acknowledge its caveats, which we discuss further in the Supplementary Information.

**Model incorporating mountain-level covariates.** To model the effect of each of these covariates, we expanded the global hierarchical model (equation (4) for raw SR and tip DR and equation (5) for sSR and tip sDR) to include mountain-level predictors:

$$\beta_m \sim \text{MVN} \left( \begin{pmatrix} \alpha_0 + \gamma_0 x_m \\ \alpha_1 + \gamma_1 x_m \\ \alpha_2 + \gamma_2 x_m \end{pmatrix}, \Sigma \right) \quad (6)$$

$$\Sigma = \begin{pmatrix} \sigma_0^2 & \dots & \rho\sigma_0\sigma_2 \\ \vdots & \ddots & \vdots \\ \rho\sigma_0\sigma_2 & \dots & \sigma_2^2 \end{pmatrix}.$$

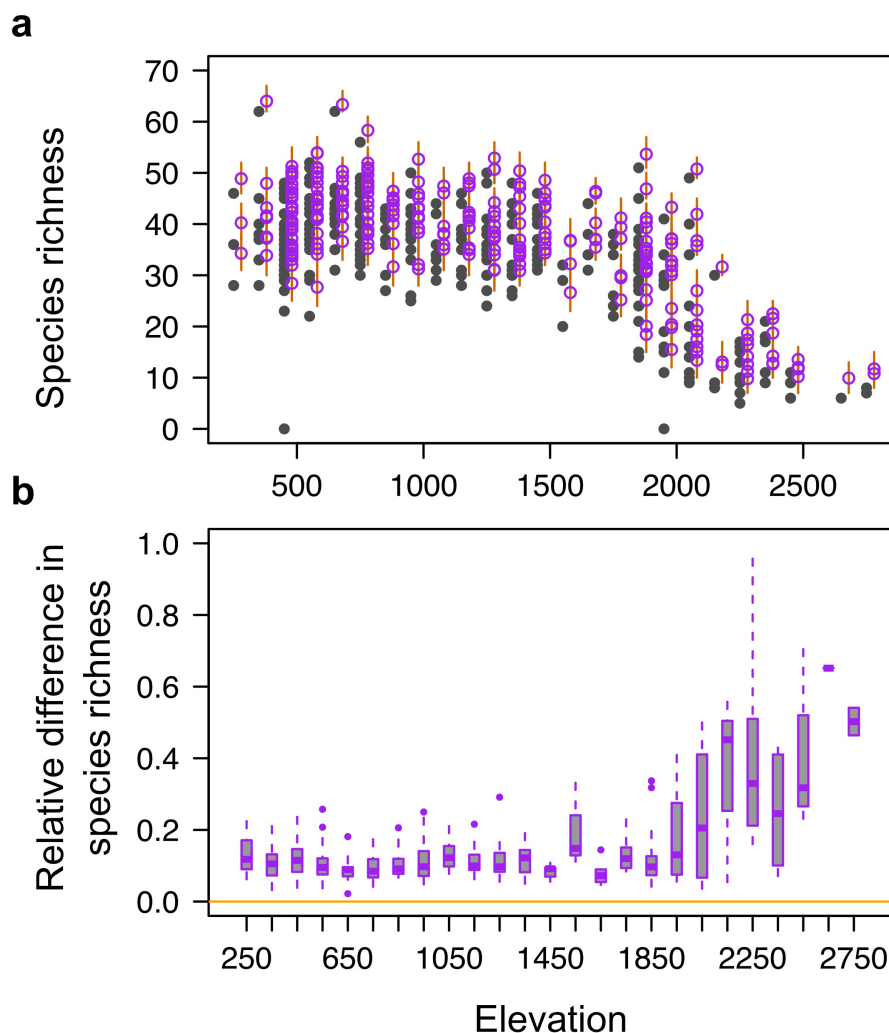
Here, the regression parameters for mountain  $m$  come from a multivariate normal distribution that is influenced by covariates  $x_m$ . The effect of each mountain-level predictor is given by  $\gamma$ , in which  $\gamma = \{\gamma_0, \gamma_1, \gamma_2\}$  is the effect on the intercept, linear and quadratic coefficient, respectively. For the effects to be comparable, we standardized the predictors by rescaling the distribution to have mean of 0 and s.d. of 1. For the subsampling analyses, we reduced the number of subsampling iterations from 100 to 20 iterations because of computational limitations (for example, with 50 iterations the model uses more than 500 GB of RAM). All models were run using R-INLA<sup>46</sup>.

Finally, given that tip DR has a phylogenetic signal, we used simulations to explore whether this non-independence could extend to the assemblage level and thereby affect effective sample size or parameter estimates of assemblage-tip sDR relationships. While we found somewhat inflated type I error at an individual mountain system level, global results do not suffer from an increased false positive rate and in neither single mountain systems nor global analyses parameter estimates are biased (see Supplementary Information).

**Data availability.** All new datasets generated during this study are included as Supplementary Information.

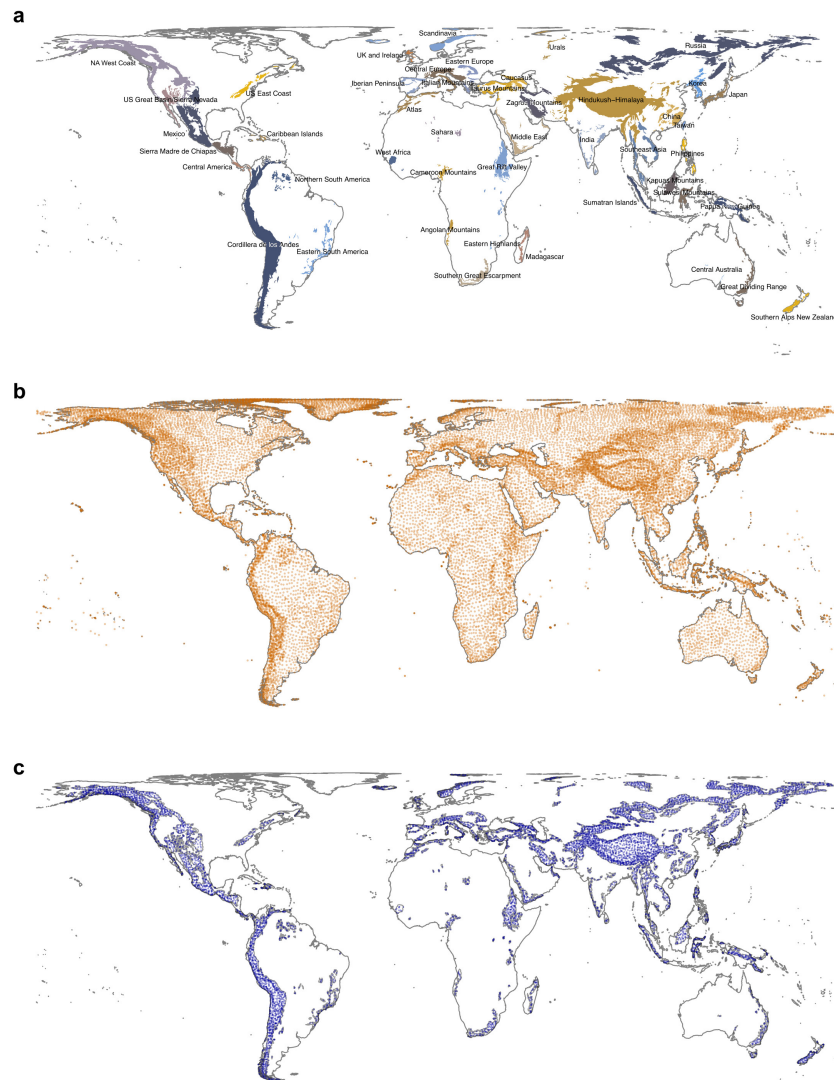
33. Royle, J. A. & Dorazio, R. Parameter-expanded data augmentation for Bayesian analysis of capture–recapture models. *J. Ornithol.* **152**, 521–537 (2012).
34. Kéry, M. & Royle, J. A. *Applied Hierarchical Models in Ecology* (Academic, 2015).
35. Plummer, M. JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling (2003).
36. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015).
37. Legendre, P. & Legendre, L. *Numerical Ecology* (Elsevier, 2012).
38. Robinson, N., Regetz, J. & Guralnick, R. P. EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. *ISPRS J. Photogramm. Remote Sens.* **87**, 57–67 (2014).
39. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
40. Bodenhofer, U., Kothmeier, A. & Palme, J. *aplcluster: Affinity Propagation Clustering* (2013).
41. Jetz, W., Wilcove, D. S. & Dobson, A. P. Projected impacts of climate and land-use change on the global diversity of birds. *PLoS Biol.* **5**, e157 (2007).
42. del Hoyo, J., Elliott, A., Sargatal, J., Christie, D. A. & de Juana, E. (eds.) *Handbook of the Birds of the World Alive* (Lynx Edicions, 2015).
43. Cleveland, W. S., Grosse, E. & Shyu, W. M. in *Statistical Models in S* (eds. Chambers, J. M. & Hastie, T. J.) (Wadsworth & Brooks/Cole, 1992).
44. Jetz, W., McPherson, J. M. & Guralnick, R. P. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* **27**, 151–159 (2012).
45. Walther, B. A. & Martin, J.-L. Species richness estimation of bird communities: how to control for sampling effort? *Ibis* **143**, 413–419 (2001).
46. Martins, T. G., Simpson, D., Lindgren, F. & Rue, H. Bayesian computing with INLA: New features. *Comput. Stat. Data Anal.* **67**, 68–83 (2013).
47. Hijmans, R. J. & van Etten, J. *raster: Geographic Analysis and Modeling with raster. Data (Kb.)* (2012).
48. Bivand, R., Keitt, T. & Rowlingson, B. *rgdal: Bindings for the Geospatial Data Abstraction Library* (2015).
49. Plummer, M. & Stukalov, A. *rjags: Bayesian Graphical Models using MCMC* (2013).
50. Gelman, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24**, 997–1016 (2014).
51. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Series B Stat. Methodol.* **71**, 319–392 (2009).
52. Huelsenbeck, J. P., Rannala, B. & Masly, J. P. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* **288**, 2349–2350 (2000).
53. Rabosky, D. L. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS One* **9**, e89543 (2014).
54. Running, S. W. *et al.* A continuous satellite-derived measure of global terrestrial primary production. *Bioscience* **54**, 547–560 (2004).
55. New, M., Lister, D., Hulme, M. & Makin, I. A high-resolution data set of surface climate over global land areas. *Clim. Res.* **21**, 1–25 (2002).
56. Prum, R. O. *et al.* A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**, 569–573 (2015).
57. Otto-Bliesner, B. L., Marshall, S. J., Overpeck, J. T., Miller, G. H. & Hu, A. Simulating Arctic climate warmth and icefield retreat in the last interglaciation. *Science* **311**, 1751–1753 (2006).
58. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
59. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).





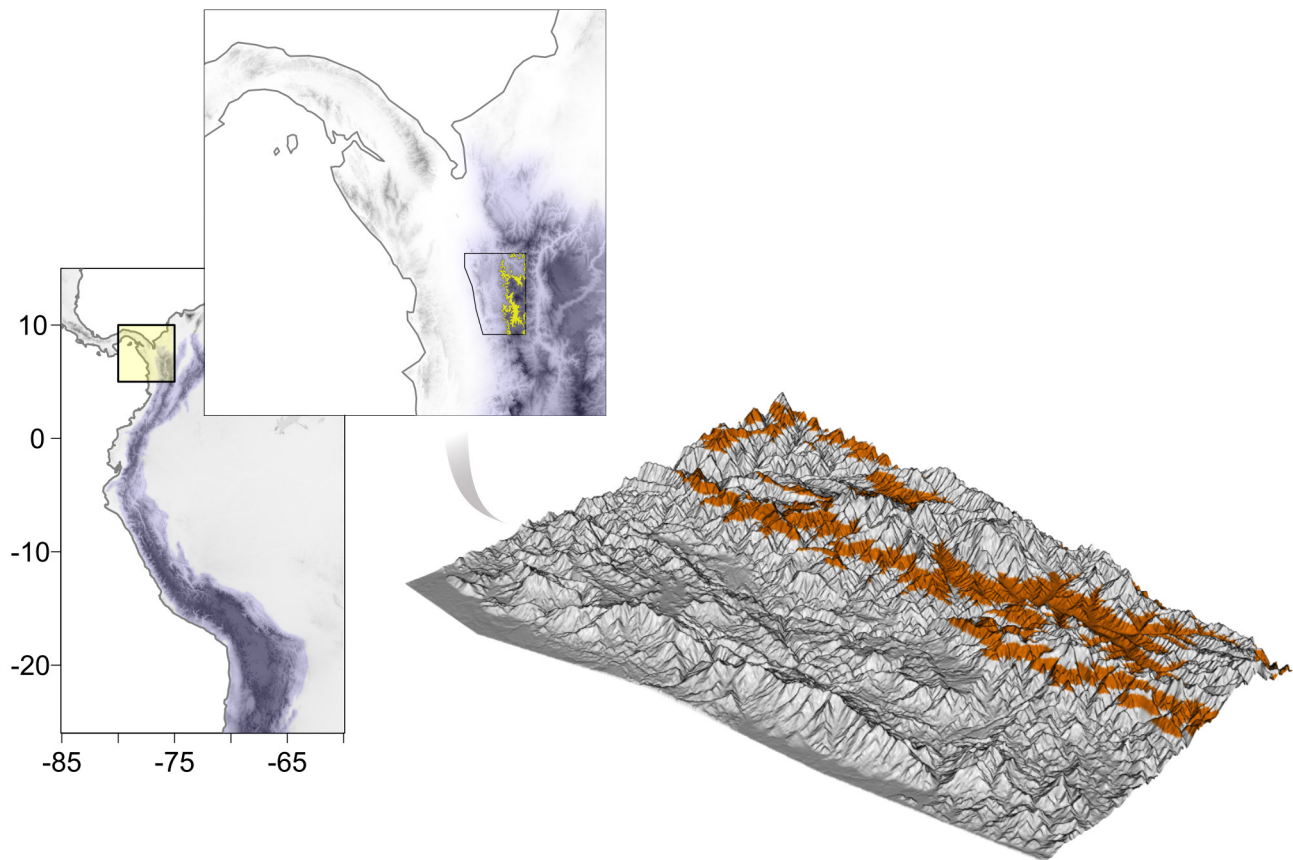
**Extended Data Figure 1 | Elevational biases in species detection in Switzerland.** Results from *N*-mixture occupancy model showcasing the differences in species richness between the observed naive species richness, and the modelled species richness. **a**, Naive and estimated species richness at each of the 267 locations along elevation ( $n = 41,652$  species presence/absence observations at different times). Closed grey points correspond to the observed richness while open blue points and orange bars correspond to the posterior average and 95% CI, respectively, of the estimated richness. To reduce figure cluttering, the estimated values are to the right of the corresponding elevation. **b**, Differences between naive and estimated richness relative to the observed species richness of the

locality every 250 m ( $n = 267$  localities). Effectively, this is the estimated proportion of species that are present but were not detected during the survey. As corroborated in the model parameters (Supplementary Information), the probability of detection of species decreases with higher elevations. Boxplot centres corresponds to the median, the upper and lower limit of the boxes correspond to the interquartile range, whiskers represent maximum or minimum points that do not exceed  $1.5 \times$  the interquartile distance, and outliers correspond to points that exceed this distance. From left to right the  $n$  for each boxplot is: 3, 8, 33, 24, 15, 20, 11, 12, 7, 11, 15, 13, 11, 4, 6, 6, 19, 13, 12, 3, 8, 6, 4, 1, 2.



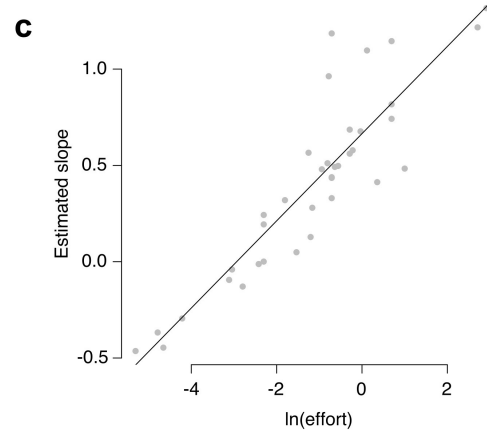
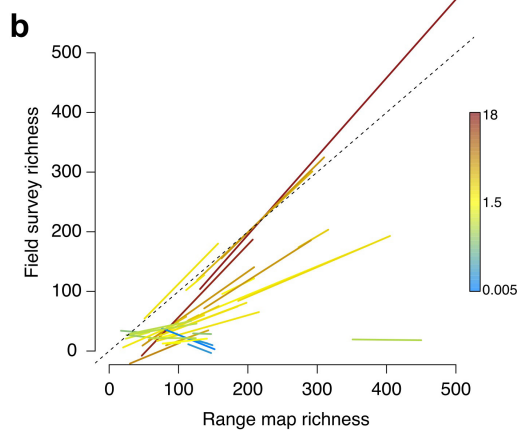
**Extended Data Figure 2 | Global mountain systems and sampling locations.** **a**, The final 46 distinct global mountain systems used in this study. **b**, Map of the sampling locations used in this study across the globe. Each sampling location is at least 1° apart in longitude and latitude and 500 m apart in elevation from every other sampling unit. This sampling

allows for denser sampling in slopes and thus takes advantage of the added resolution brought by the species' elevational ranges. **c**, Map of the subset of sampling locations that lie within the mountain systems in this study. Continental coastlines from Natural Earth.



**Extended Data Figure 3 | Sampling unit.** Graphical representation of a sampling unit used in this study. We tailored our sampling to the resolution of our data: range maps validated to approximately 1° latitude/longitude and 500 m elevation. The yellow coloured area in the map inset

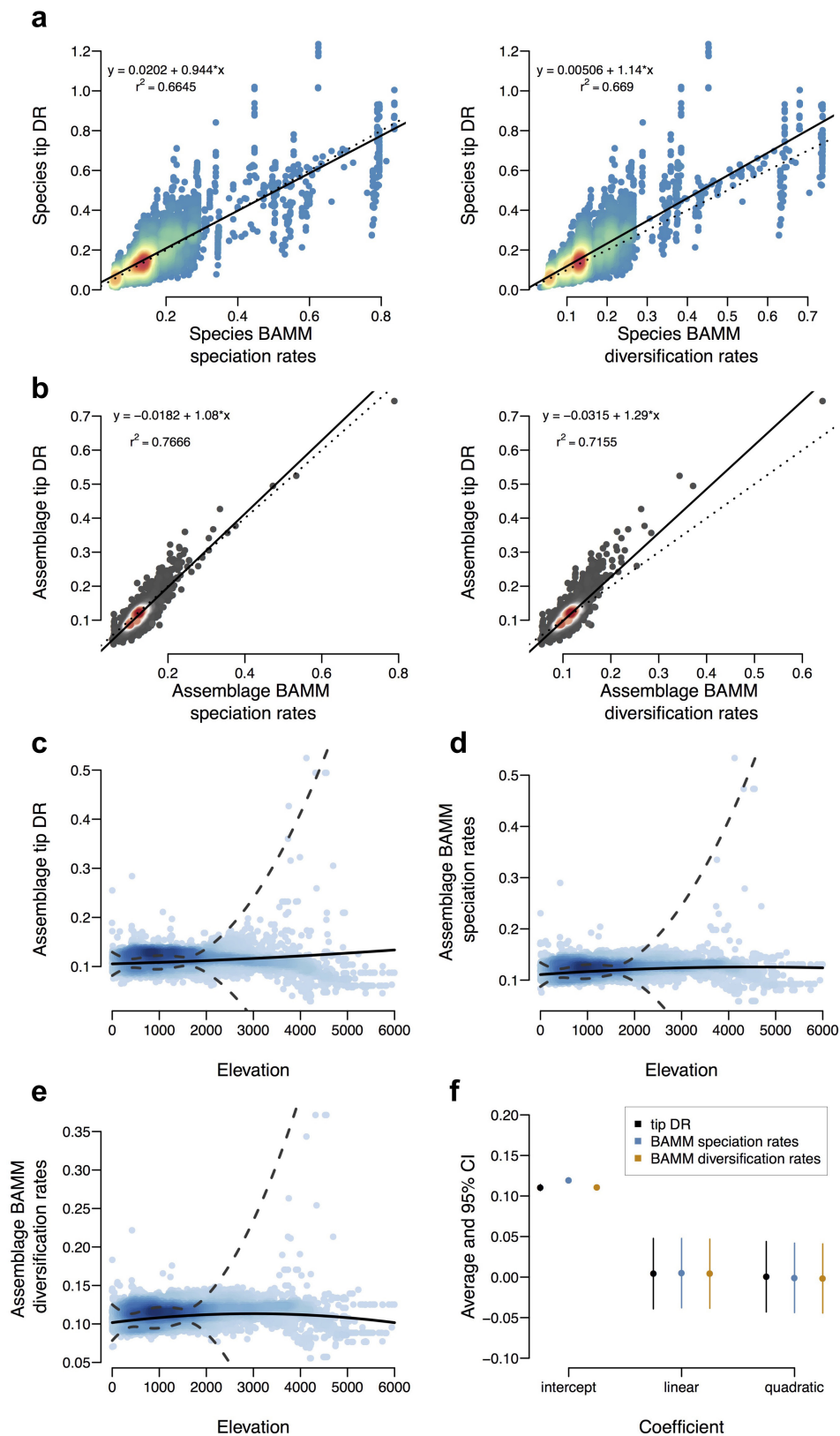
corresponds to the orange coloured area in the 3D plot and represents one such sampling unit at a random location in the Colombian Andes. Continental coastlines from Natural Earth.



**Extended Data Figure 4 | Comparisons between field work and range map species inventories.** Data and results from the multilevel regression between richness estimates from map ranges and field surveys. **a**, Each panel corresponds to an elevational transect in which species richness inventories were conducted. Red points correspond to the accrued richness count in the field and green points correspond to the richness inferred from expert range maps and elevational ranges at the same geographical coordinates. Each fieldwork locality varies markedly in

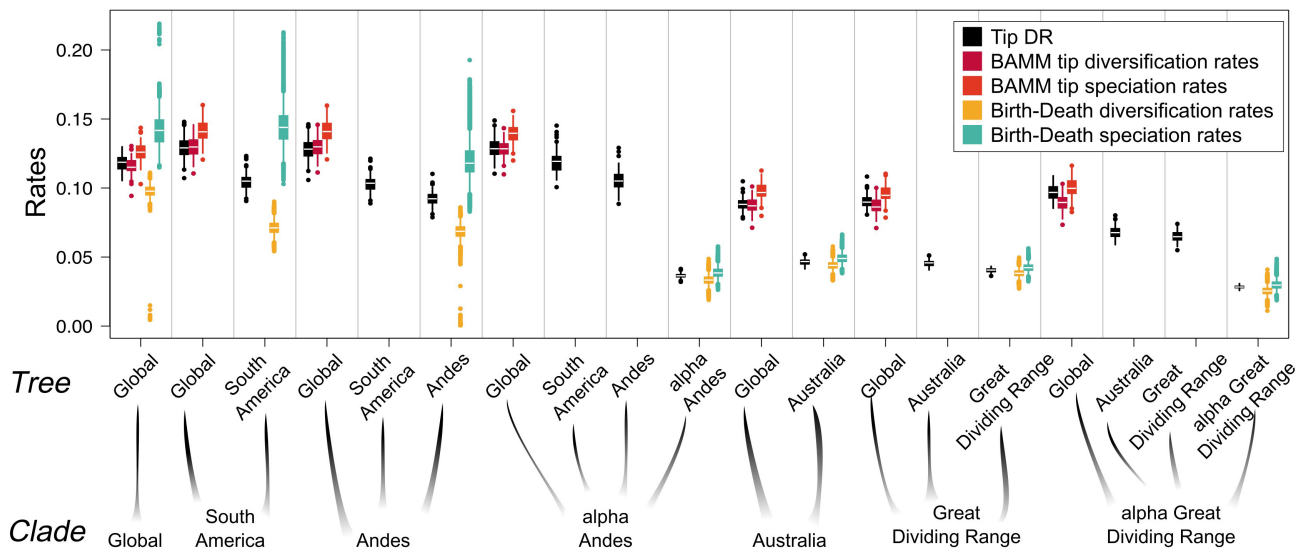
sampling effort, focal group and elevational extent, among others; for specific information on each transect see Supplementary Table 2. **b**, Comparison of richness estimates: each line corresponds to the association between richness estimates from a field survey and expert range maps. The colour warmth corresponds to the natural logarithm of sampling effort. The dashed line corresponds to the 1:1 line. **c**, Plot of the estimated slope for each survey versus range map richness association against the natural logarithm of sampling effort.





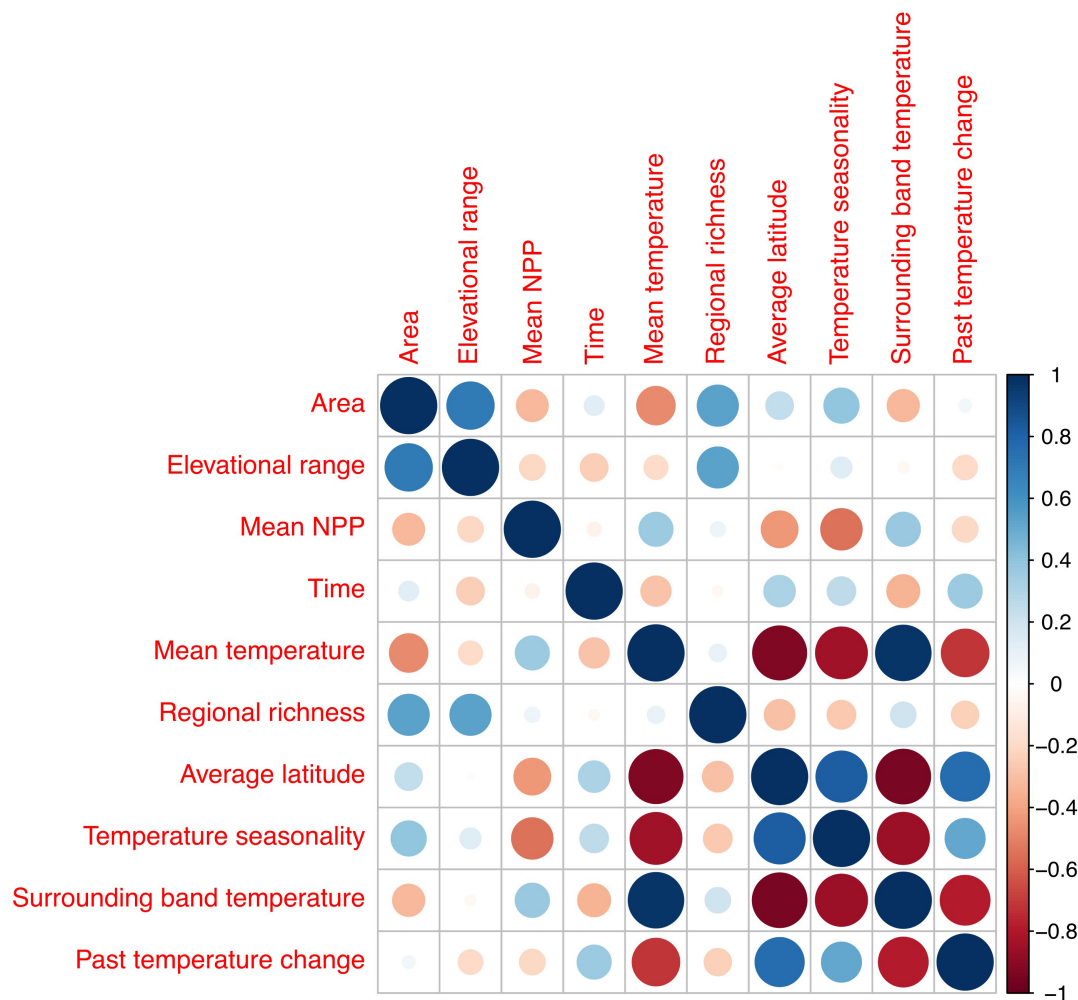
**Extended Data Figure 5 | Comparison between tip DR and BMM tip rates.** **a**, Average species tip DR versus BMM tip speciation (left) and diversification rates (right). Continuous line corresponds to an ordinary linear regression; dashed line corresponds to the 1:1 line ( $n = 9,993$  extant species). **b**, Average assemblage tip DR versus BMM tip speciation (left) and diversification (right) rates. Continuous line corresponds to an ordinary linear regression; dashed line corresponds to the 1:1 line. See Supplementary Information for further information. **c–e**, Comparison

between multilevel models of average assemblage tip DR (**c**), BMM tip diversification (**d**) and speciation rates (**e**) along elevation (equation (4) in Methods). In each regression, the black line corresponds to the mean while the grey dashed lines correspond to the 95% credible interval (CI). Higher colour intensity corresponds to higher density of points. **f**, Comparison between the intercept and the linear and quadratic coefficients (posterior average and 95% CI) from the three regressions in **c–e**. For **b–f**,  $n = 8,410$  assemblages.

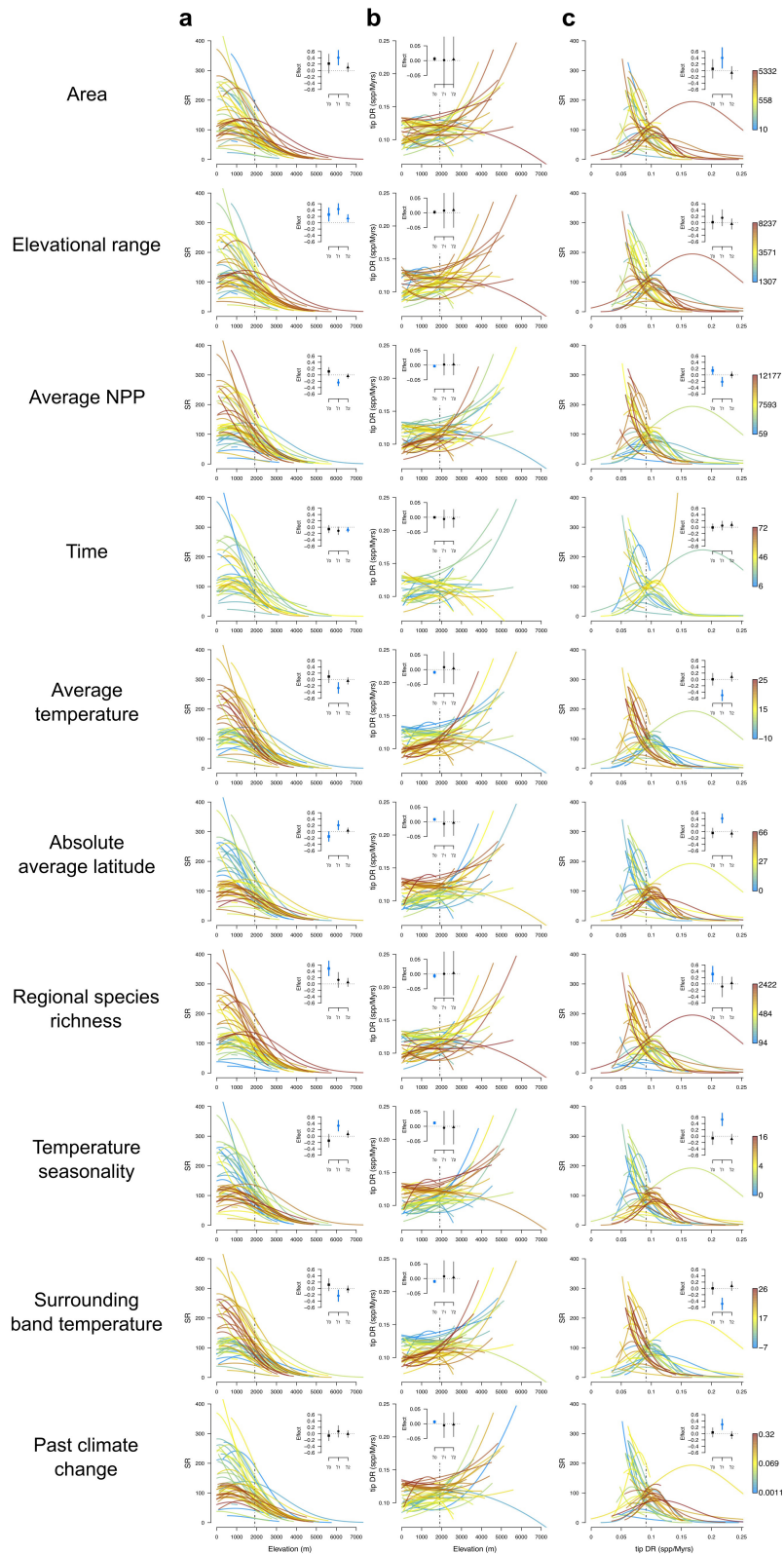


**Extended Data Figure 6 | Comparison between tip DR and clade-level diversification metrics.** Comparison between average tip DR, estimates of speciation and diversification given a birth–death model, and average BAMM tip diversification and speciation rates for clades defined through different nested spatial boundaries and different trees. ‘Tree’ refers to the species group present in the tree when estimating the rates, while ‘clade’ refers to the species group for which rates are being estimated. For instance, if the clade is ‘Andes’ and the tree is ‘South America’, then we used a tree with every non-South American species pruned out, and estimated the rates among the species present in the Andes only. Global corresponds to all bird species, South America to all South American species, Andes to all species present in our mountain delineation of the Cordillera de los Andes, and alpha Andes to all species intersecting with a random point

within the Cordillera de los Andes. Similarly, Australia corresponds to all Australian species, Great Dividing Range to all species present in our delineation of the Great Dividing Range and alpha Great Dividing Range to all species intersecting with a random point within this mountain system. Boxplot centres show to the median, the upper and lower limit of the boxes correspond to the interquartile range, whiskers represent maximum or minimum points that do not exceed  $1.5\times$  the interquartile distance, and outliers correspond to points that exceed this distance. Each boxplot consists of: for tip DR,  $n = 100$  tip DR harmonic means for 100 trees; for BAMM diversification and speciation rates,  $n =$  harmonic averages across the species for 50 trees; for birth–death diversification and speciation rates,  $n = 10,000$  samples from the posterior.



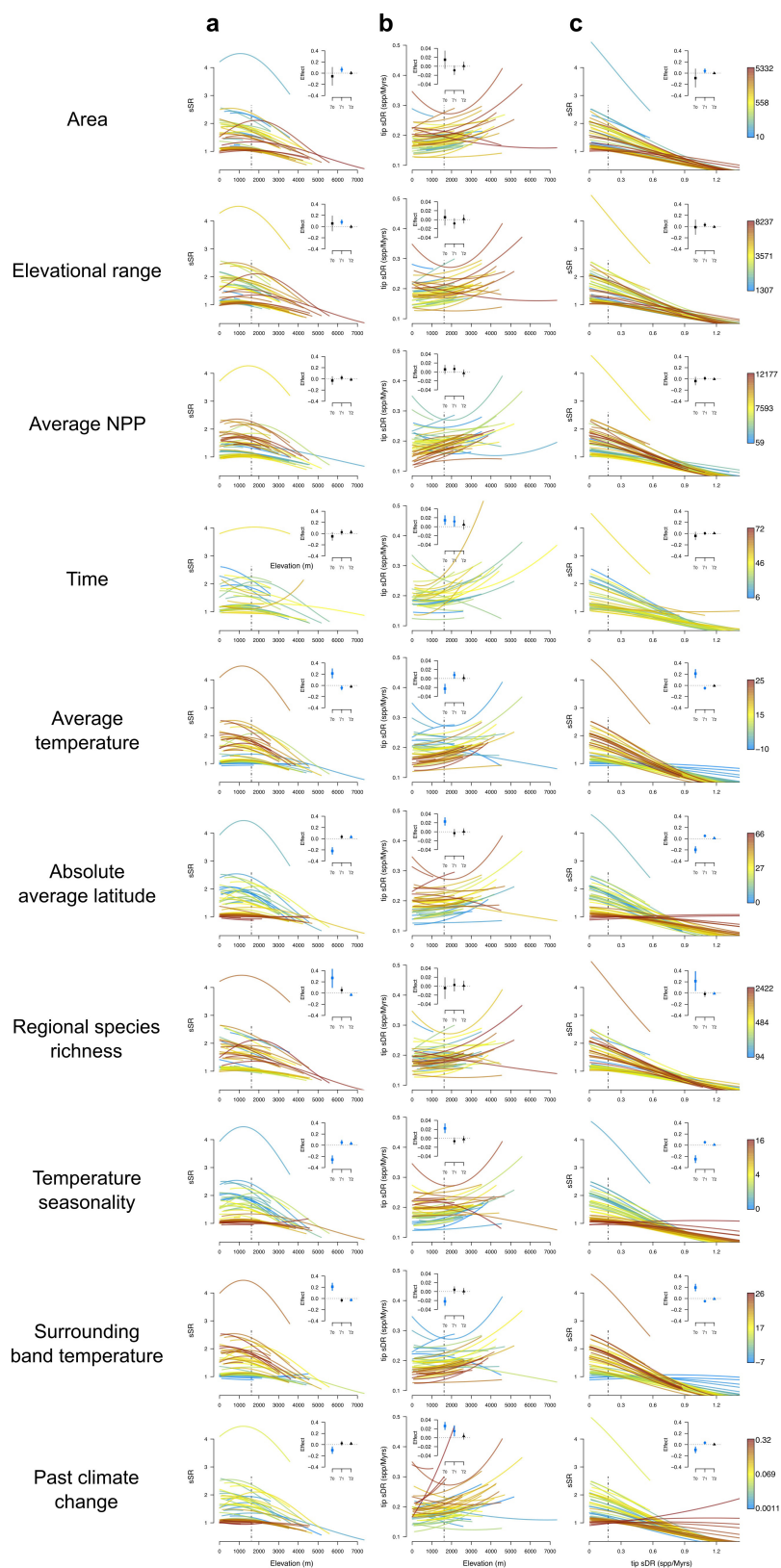
**Extended Data Figure 7 |** Correlation matrix between the different mountain system level predictors used. Pearson's correlation coefficient for mountain level covariates ( $n = 46$ ).



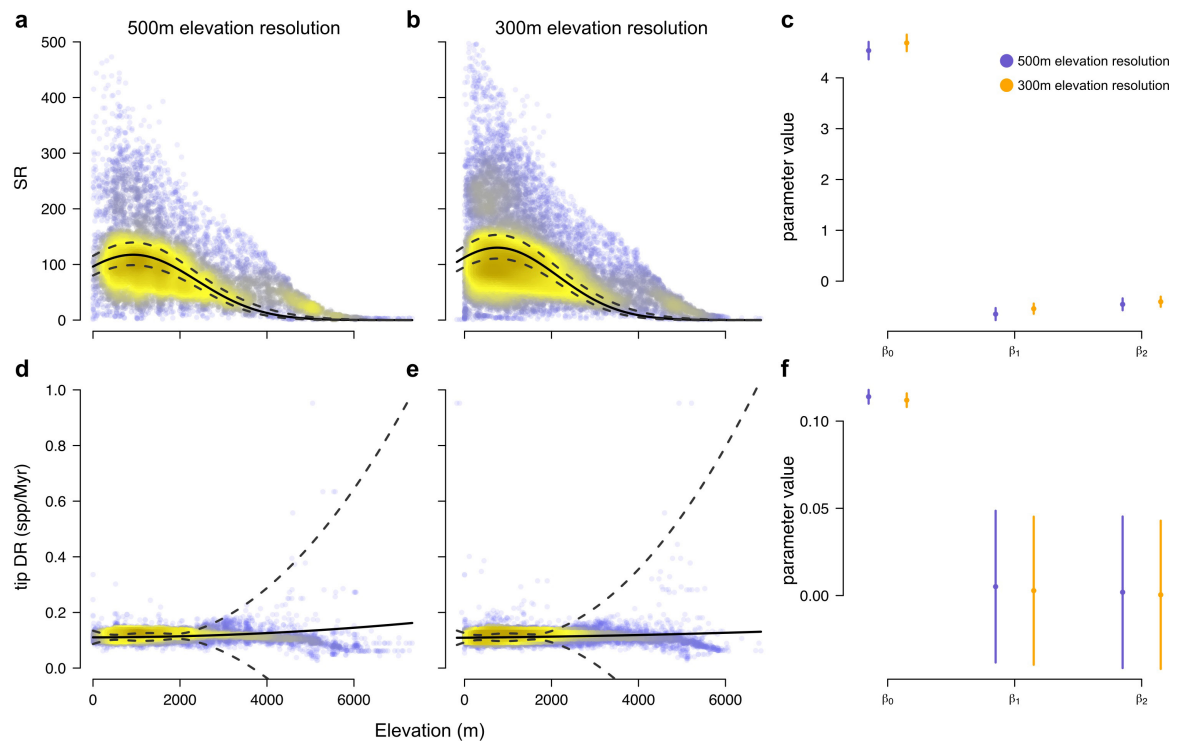
**Extended Data Figure 8 | Mountain system level predictors for non-subsampled data.** Global multilevel models with mountain system as random effect and mountain-level predictors for non-subsampled assemblages. We used total mountain area, total elevational range, average NPP, average temperature, mountain system age, average latitude, average temperature seasonality, average surrounding band temperature, regional species richness and past rates of temperature change. Each line corresponds to a mountain system; lines are coloured such that the redder end of the spectrum corresponds to higher covariate values while bluer colours correspond to lower values. The inset plots show the effect of each

covariate on the intercept ( $\gamma_0$ ), the linear ( $\gamma_1$ ) and the quadratic coefficient ( $\gamma_2$ ; posterior average and 95% CI); blue coloured effects correspond to coefficients where the 95% CI does not overlap with 0 (that is,  $\text{Pr}(\text{effect is not } 0) > 0.975$ ). Vertical dashed black line corresponds to the mean of the  $x$ -axis values (that is, corresponds to '0' when the  $x$  axis is standardized). For interpretation of these results see Supplementary Information. Multilevel models assessing the effects of mountain characteristics on assemblage richness along elevation (a), assemblage tip DR along elevation (b) and the relationship of assemblage richness with tip DR (c). For all regressions,  $n = 8,410$  assemblages.





**Extended Data Figure 9 | Mountain system level predictors for subsampled data.** Global multilevel models with mountain system and sampling locality as random effect and mountain-level predictors. Figure description as in Extended Data Fig. 8. For all regressions,  $n = 42,526$  subsampled assemblages.



**Extended Data Figure 10 | Sensitivity of multilevel model results to resolution of species elevational range data.** **a, b,** Raw assemblage richness along elevation contrasting the use of 500 m (**a**) and 300 m (**b**) resolution in elevation sampling. The black solid line corresponds to the expectation of species richness along elevation, while the grey dashed

lines correspond to the 95% CI. Higher colour intensity corresponds to higher density of points. **c,** Parameter quantiles (0.025, 0.5 & 0.975) for the intercept, slope and quadratic coefficient of the regressions shown in **a, b**. **d–f,** The same relationships as **a–c** but for tip DR as response. For 500 m,  $n = 8,410$  assemblages; for 300 m,  $n = 14,218$  assemblages.

# Regeneration of the lung alveolus by an evolutionarily conserved epithelial progenitor

William J. Zacharias<sup>1,2\*</sup>, David B. Frank<sup>2,3,4\*</sup>, Jarod A. Zepp<sup>1,2</sup>, Michael P. Morley<sup>1,2,4</sup>, Farrah A. Alkhaleel<sup>1,2</sup>, Jun Kong<sup>1,2</sup>, Su Zhou<sup>1,4</sup>, Edward Cantu<sup>5</sup> & Edward E. Morrisey<sup>1,2,4,6,7</sup>

**Functional tissue regeneration is required for the restoration of normal organ homeostasis after severe injury. Some organs, such as the intestine, harbour active stem cells throughout homeostasis and regeneration<sup>1</sup>; more quiescent organs, such as the lung, often contain facultative progenitor cells that are recruited after injury to participate in regeneration<sup>2,3</sup>. Here we show that a Wnt-responsive alveolar epithelial progenitor (AEP) lineage within the alveolar type 2 cell population acts as a major facultative progenitor cell in the distal lung. AEPs are a stable lineage during alveolar homeostasis but expand rapidly to regenerate a large proportion of the alveolar epithelium after acute lung injury. AEPs exhibit a distinct transcriptome, epigenome and functional phenotype and respond specifically to Wnt and Fgf signalling. In contrast to other proposed lung progenitor cells, human AEPs can be directly isolated by expression of the conserved cell surface marker TM4SF1, and act as functional human alveolar epithelial progenitor cells in 3D organoids. Our results identify the AEP lineage as an evolutionarily conserved alveolar progenitor that represents a new target for human lung regeneration strategies.**

Wnt signalling, which is revealed by expression of the direct target gene *Axin2*, has previously been shown to have an important role in the development of both the surfactant-producing alveolar type 2 (AT2) cells and the alveolar type 1 (AT1) cells that form the gas-exchange surface of the lung alveolus<sup>4</sup>. In the lungs of adult *Axin2<sup>creERT2-tdT</sup>; R26R<sup>eYFP</sup>* (*R26R<sup>eYFP</sup>* is also known as *Gt(ROSA)26Sor<sup>tm1(eYFP)Cos</sup>*) mice, *Axin2<sup>+</sup>* Wnt-responsive epithelial cells are restricted to the alveolar region and express the AT2 cell marker *Sftpc* (Fig. 1a–d and Extended Data Fig. 1a–e). Few *Axin2<sup>+</sup>* cells express AT1 markers, including *Hopx* (Fig. 1e and Extended Data Fig. 1k, l). These *Axin2<sup>+</sup>* AT2 cells represent an alveolar epithelial progenitor lineage—hereafter referred to as AEPs—that comprises approximately 20% of adult AT2 cells (Fig. 1f). AEPs express the same level of AT2 marker genes as other AT2 cells (Extended Data Fig. 1f), and also show enriched expression of Wnt targets (Extended Data Fig. 1g). We performed one-, three- and nine-month lineage tracing using *Axin2<sup>creERT2-tdT</sup>; R26R<sup>eYFP</sup>* mice to define AEP dynamics during adult homeostasis (Fig. 1a). AEPs are notably stable, and we observed only a small increase in the number of AEP-marked cells after nine months (Fig. 1g and Extended Data Fig. 2a–c). In contrast to developmental alveologenesis<sup>4</sup> (Extended Data Fig. 3), few *Axin2<sup>+</sup>* AT2 cells become AEPs during homeostasis (Fig. 1h).

To assess dynamics of AEPs in lung injury, we used the H1N1 influenza virus to injure the lungs of adult *Axin2<sup>creERT2-tdT</sup>; R26R<sup>eYFP</sup>* mice, causing spatially heterogeneous injury that is similar to human influenza infection<sup>5</sup>. We defined four regions of injury severity: (i) zone 1, no morphological changes; (ii) zone 2, a minor injury with mild interstitial thickening; (iii), zone 3, substantial injury; and (iv) zone 4, total

alveolar destruction (Fig. 1i). We used this spatially specific response to analyse the contribution of AEPs to lung regeneration.

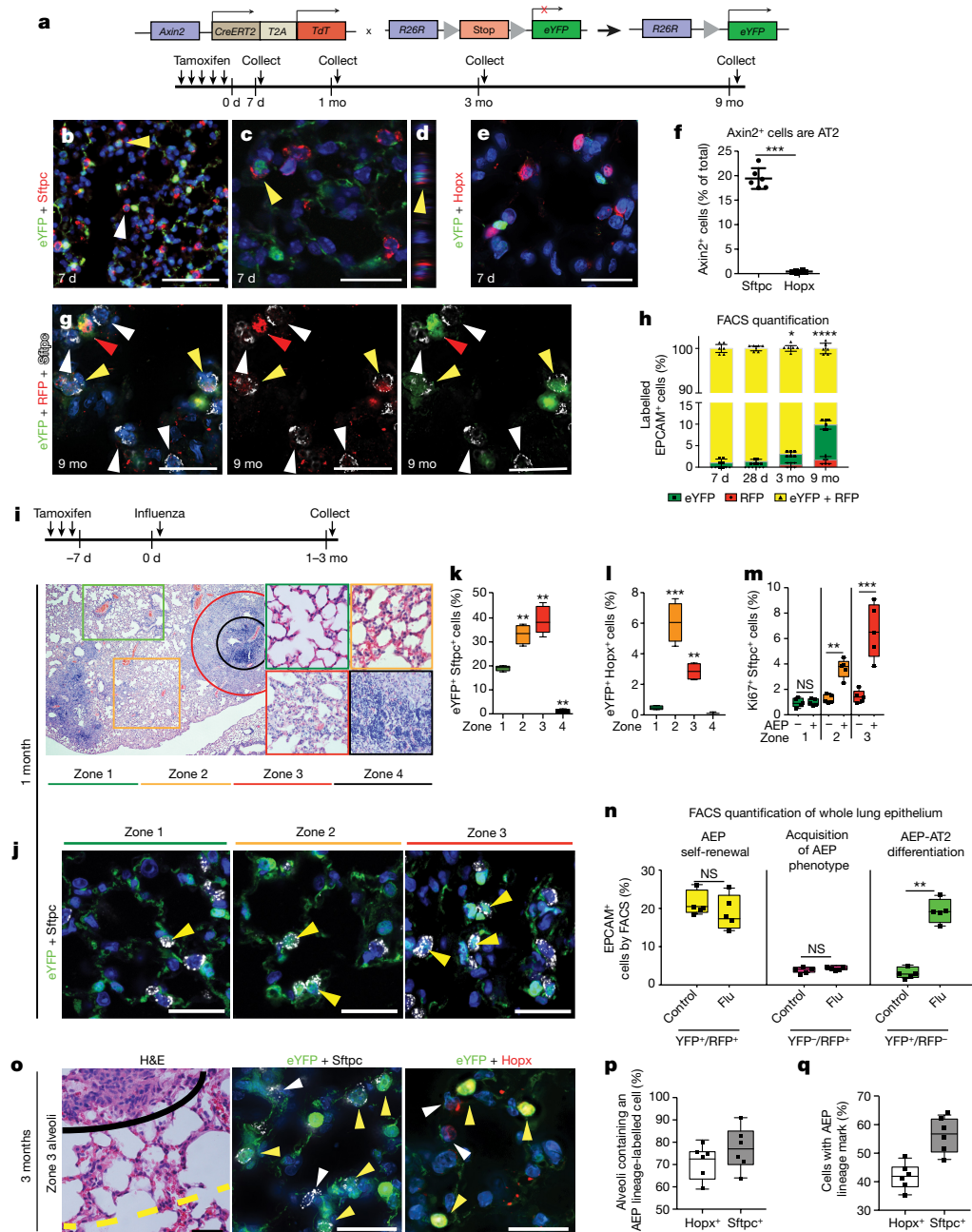
Recent studies have shown that Sox2-derived *Krt5<sup>+</sup>* epithelial cells migrate to damaged distal lung regions to re-create an epithelial barrier<sup>6–10</sup>. We observe *Krt5<sup>+</sup>* epithelium specifically in zone 4 after influenza infection (Extended Data Fig. 4a–d, f), but lineage tracing demonstrates that no *Krt5<sup>+</sup>* cells are derived from AEPs (Extended Data Fig. 4g). Furthermore, AEPs express minimal levels of *Krt5* or *Sox2* RNA and no detectable protein (Extended Data Figs 1f, 4e), further indicating that AEPs and *Krt5<sup>+</sup>* cells derive from distinct lineages. In zone 4, *Sftpc<sup>+</sup>* and *Krt5<sup>+</sup>Sftpc<sup>+</sup>* cells are very rare (Extended Data Fig. 4l), which confirms previous reports<sup>7</sup> that the *Krt5<sup>+</sup>* lineage cells do not efficiently regenerate *Sftpc<sup>+</sup>* cells except after forced Wnt activation<sup>9</sup>.

One month after influenza injury, AEPs and their progeny are present at homeostatic levels in zone 1. However, in zones 2 and 3 the number of AT2 cells expands significantly (Extended Data Fig. 4h)<sup>11,12</sup>, with a large increase in the percentage of AT1 and AT2 cells arising from the AEP lineage (Fig. 1j–l and Extended Data Figs 2d–i, 4j–l). This robust labelling is independent of the timing of tamoxifen injection before influenza infection (Extended Data Fig. 5g–h). Notably, in zone 2 and zone 3 the AEP lineage shows a marked and specific increase in proliferation (Fig. 1m and Extended Data Fig. 2k–o). Three months after injury, within 300 microns of a persistent *Krt5<sup>+</sup>* pod, a majority of AT2 cells and many AT1 cells in regenerated alveoli are derived from the AEP lineage (Fig. 1o–q). Immunohistochemistry and fluorescence-activated cell sorting (FACS) analysis after influenza injury demonstrate that AEPs self-renew to maintain the AEP lineage and generate a large number of new lineage-traced alveolar epithelial progeny (Fig. 1n and Extended Data Figs 2j, 5a–e). Notably, few non-AEP AT2 cells acquire the AEP phenotype even in the setting of considerable lung injury (Fig. 1n and Extended Data Fig. 5e).

AEPs exhibit a distinct gene expression profile enriched in lung developmental genes (Fig. 2a–d), including the key genes *Fgfr2*, *Nkx2-1*, *Id2*, *Etv4*, *Etv5* and *Foxa1* (Extended Data Fig. 6 and Supplementary Table 1). Furthermore, analysis by assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Extended Data Fig. 7) revealed a marked difference between AEPs and AT2 cells, with more than 40% of the genome containing differential open chromatin (Fig. 2a). Although many regions of common open chromatin are found near housekeeping genes, regions of AEP-enriched open chromatin are found near lung development genes (Extended Data Fig. 7c). DNA binding-site motif analysis shows that AEP-enriched chromatin contains binding sites for AEP-enriched transcription factors of the Klf, Six, Sox, Nkx2 and Elf/Ets families (Extended Data Fig. 7d, e), all of which are known to be regulators of progenitor cell behaviour<sup>13–17</sup>. Moreover, a group of primed cell-cycle regulators near AEP-enriched open chromatin were

<sup>1</sup>Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>2</sup>Penn Center for Pulmonary Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>3</sup>Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. <sup>4</sup>Penn Cardiovascular Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>5</sup>Department of Surgery, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>6</sup>Department of Cell and Developmental Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>7</sup>Penn Institute for Regenerative Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

\*These authors contributed equally to this work.



**Figure 1 | Identification of an *Axin2*<sup>+</sup> AEP in the adult lung that regenerates a substantial percentage of the alveolar epithelium.**

**a**, Schematic of *Axin2<sup>CreERT2</sup>;R26R<sup>eYFP</sup>* mice. Lineage tracing experimental design is as indicated. D, day; mo, month. **b–d**, *Axin2* marks a subset of AT2 cells. eYFP is detected by an anti-GFP antibody. Unmarked, white arrowheads. AEP-marked, yellow arrowheads; **d** represents an orthogonal view of **c**. **e**, *Hopx*<sup>+</sup> AT1 cells are not marked by eYFP. **f**, Approximately 20% of AT2 cells express *Axin2*. **g, h**, Epithelial Wnt responsiveness is stable for up to nine months. The majority of the AEP lineage remains *Axin2<sup>tdT</sup>*-positive; some AEP progeny lose *Axin2<sup>tdT</sup>* expression. Very few *Sftpc*<sup>+</sup>*Axin2*<sup>−</sup> cells gain *Axin2<sup>tdT</sup>* expression. Red arrow, an *Axin2*<sup>+</sup> mesenchymal cell. **i**, Experimental design is as indicated. Influenza-induced lung injury results in regionalized alveolar damage: minimal (zone 1), mild (zone 2), severe (zone 3) or complete (zone 4). **j–l**, AEP-generated *Sftpc*<sup>+</sup> cells (**j, k**) and *Hopx*<sup>+</sup> AT1 cells (**l**) expand in

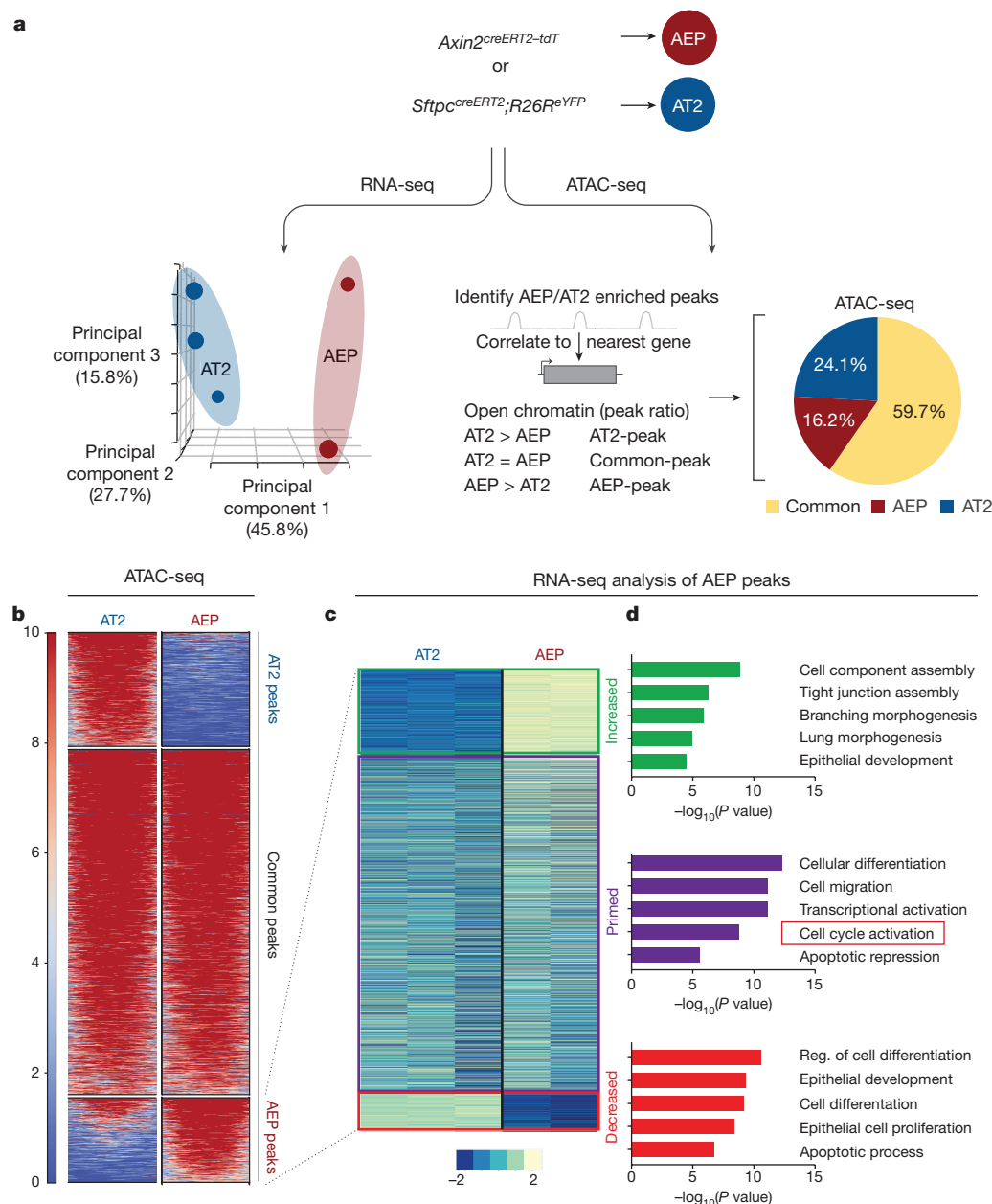
zones 2 and 3. **m**, *Ki67*<sup>+</sup> AEPs preferentially re-enter the cell cycle in areas of regeneration. **n**, AEPs can self-renew (*YFP*<sup>+</sup>*RFP*<sup>+</sup>), but very few non-AEP cells acquire *Axin2* expression (*YFP*<sup>−</sup>*RFP*<sup>+</sup>). Flu, cells from mice with influenza-induced lung injury. **o**, A region of regenerated lung epithelium near a persistent *Krt5*<sup>+</sup> pod. Black line, border of *Krt5*<sup>+</sup> pod. Yellow dotted line, region of regeneration. **p, q**, A large number of new AEP-derived AT1 and AT2 cells are found within three alveolar units (regenerated zone 3) of *Krt5*<sup>+</sup> pods. **n = 5** (**m, n**), **6** (**f–g, o, p**) or **10** (all other panels) mice from 2 (**g, h, o, p**) or 3 (all other panels) individual experiments. Statistics are representative of all biological replicates. Plots are centred on mean, with bars indicating s.d. \**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001 and \*\*\*\**P* < 0.0001 by two-tailed *t*-test (**e, p, q**) or ANOVA with adjustment for multiple comparison testing (all other panels). NS, not significant. Scale bars: **b**, 100 μm; **c–e, g, j, o**, 50 μm.

dynamically regulated in AEPs two weeks after influenza infection<sup>18–21</sup> (Fig. 2b–d and Extended Data Fig. 8a–c).

To isolate human AEPs, we identified cell surface markers that were enriched in mouse AEPs (Fig. 3a). These studies identified

the epithelial cancer stem cell membrane protein *Tm4sf1*<sup>22,23</sup> as a marker for mouse AEPs (Fig. 3b and Extended Data Fig. 8a–c). Immunohistochemistry and FACS analysis demonstrates that *Tm4sf1* marks approximately 20% of labelled mouse AT2 cells and





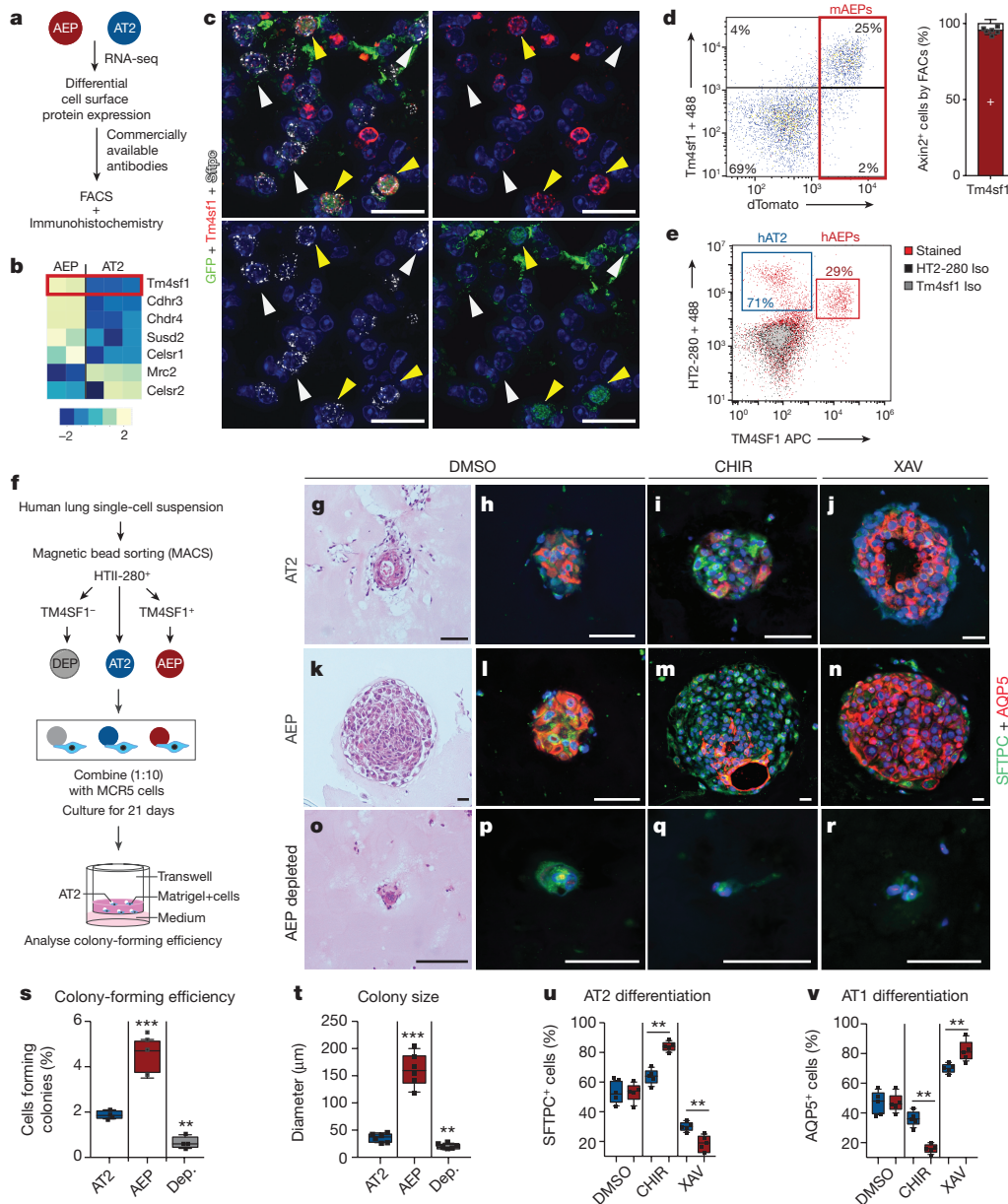
**Figure 2 | AEPs possess a distinct transcriptome and chromatin architecture enriched for cell-cycle and progenitor-cell pathways.** **a**, RNA-seq and ATAC-seq was performed on AEPs and AT2 cells. Principal component analysis plot of the top 500 most-variable genes, showing that AEPs segregate into a distinct population from AT2 cells. ATAC-seq shows that more than 40% of the genome was differentially accessible in AEPs (16.2%) or AT2 cells (24.1%). **b**, ATAC-seq heat map showing genome-wide regions of differential open chromatin peaks in AT2 cells versus AEPs. **c**, AEP-enriched ATAC-seq peaks compared to RNA-seq expression shows that a majority of genes associated with AEP open chromatin are not differentially expressed, but are primed for rapid activation. **d**, A subset of these primed genes are associated with cell-cycle activation. Reg., regulation. For full details of the experimental design and statistical methods for these analyses, see Methods.

more than 90% of mouse AEPs (Fig. 3c, d and Extended Data Fig. 8a). Using a combination of a human TM4SF1 antibody (Extended Data Fig. 8d) and human-AT2-specific HTII-280 antibody<sup>24</sup> (Extended Data Fig. 8b, e–h), we were able to identify a distinct subset of HTII-280<sup>+</sup>TM4SF1<sup>+</sup>EPCAM<sup>+</sup> putative human AEPs in normal human lung. These human AEPs comprise approximately 29% of the human AT2 population (Fig. 3e) and express *SFTPC*, but not *KRT5* or *SOX2*, mRNA (Supplementary Table 2).

Using clonal alveolar organoid assays<sup>25</sup>, both mouse AEPs and human AEPs form a greater number of, and larger, organoids that contain both AT1 and AT2 cells but no *SOX2*<sup>+</sup> or *KRT5*<sup>+</sup> cells (Extended Data Fig. 8i, j). AEPs also demonstrate increased responsiveness to Wnt modulation when compared to AT2 cells (Fig. 3f–n and Extended Data Fig. 9). Notably, depletion of TM4SF1<sup>+</sup> cells from the human AT2 population leads to a dramatic loss of organoid formation (Fig. 3o–s). Wnt inhibition promoted AT1 cell differentiation and Wnt activation promoted AT2 cell formation in both mouse and human organoids, but not in human AEP-depleted organoids (Fig. 3o–r, u, v and Extended Data Fig. 9o, p). These data suggest that TM4SF1<sup>+</sup>HTII-280<sup>+</sup> human AEPs are the functional equivalent of mouse AEPs.

RNA sequencing analysis (RNA-seq) demonstrated that a large proportion of human AEP-enriched genes (35.6%)—including key progenitor cell regulators—were evolutionarily conserved with mouse AEPs (Fig. 4a, b and Extended Data Fig. 10a, b). In particular, mouse and human AEPs are both enriched for Wnt pathway targets, including *AXIN2* and *FGFR2*; *FGFR2* is the primary receptor for FGF7 and FGF10<sup>26–32</sup> (Fig. 4c, Extended Data Fig. 10k and Supplementary Table 2). DNA binding motif analysis shows that TCF/LEF binding sites are enriched in open chromatin near conserved AEP genes, and  $\beta$ -catenin is bound to some of these genomic regions (Extended Data Fig. 10c–e), supporting the idea that the Wnt-responsiveness of AEPs is evolutionarily conserved. Treatment of both mouse and human AEPs with FGF7 or FGF10 ligand resulted in substantial increases in colony size and colony-forming efficiency, whereas mouse AT2 and AEP-depleted human AT2 cells exhibited a diminished response (Fig. 4d–w and Extended Data Fig. 10f–q).

Our data reveal that AEPs are a major lineage that contributes to functional alveolar epithelial regeneration by producing a large number of both AT2 and AT1 cells after injury. By contrast, Sox2-derived



**Figure 3 | Identification of Tm4sf1 as an AEP-specific cell surface marker capable of isolating functional human AEPs.** **a**, **b**, AEP-enriched cell-surface proteins with an available antibody. Tm4sf1 is highlighted. **c**, Tm4sf1 is expressed in mouse AEPs. Yellow arrow, AEP. White arrow, AT2 cell. **d**, FACS analysis ( $n = 6$  individual mice, see Extended Data Fig. 8) demonstrating that Tm4sf1 correlates strongly with Axin2<sup>tdT</sup> expression. **e**, A human anti-TM4SF1 antibody marks a subset of human HTII-280<sup>+</sup> AT2 (hAT2) cells, which are putative human AEPs (hAEPs). 488, Alexa Fluor 488. **f**, Diagram of human lung alveolar organoid assay using either total human AT2 cells (HTII-280<sup>+</sup>), human AEPs (HTII-280<sup>+</sup>TM4SF1<sup>+</sup>) or AEP-depleted human AT2 cells (HTII-280<sup>+</sup>TM4SF1<sup>-</sup>). Indicated cultures were treated with CHIR or XAV to modulate Wnt signalling.

**g–j**, The complete human AT2 population generates alveolar organoids and responds to Wnt activation by increasing AT2 cell differentiation, or to Wnt inhibition by increasing AT1 cell differentiation. **k–n**, Human AEPs generate a greater number of, and larger, organoids that respond more robustly to Wnt modulation. **o–r**, Depletion (Dep.) of TM4SF1<sup>+</sup> cells from human AT2 cells results in a loss of alveolar organoid formation and Wnt responsiveness. **s–v**, Quantification of colony forming efficiency (**s**), colony size (**t**), AT2 (**u**) and AT1 (**v**) cell differentiation.  $n = 4$  individual human organoid experiments. Statistics are inclusive of all biological replicates. \*\* $P < 0.01$  and \*\*\* $P < 0.001$  by ANOVA with adjustment for multiple comparison testing. Plots are centred on mean with bars indicating standard deviation. Scale bars: **c**, 50 μm; **f–q**, 25 μm.

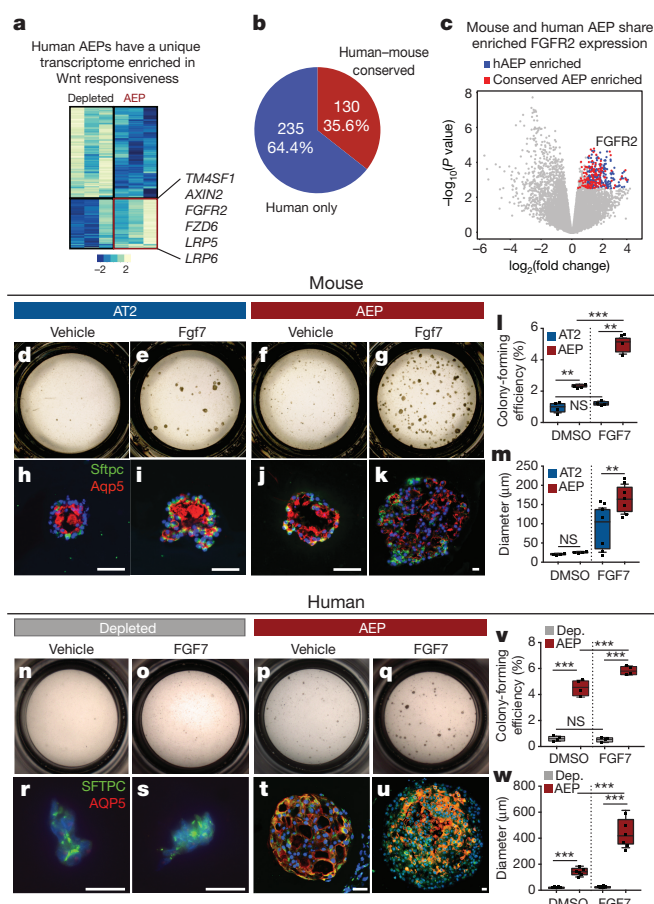
Krt5<sup>+</sup> cells migrate from the proximal airway after acute lung injury, preventing loss of the epithelial barrier<sup>6–9</sup>. AEPs and Krt5<sup>+</sup> cells are likely to act in concert: Krt5<sup>+</sup> cells probably act rapidly to prevent immediate loss of epithelial barrier while AEPs simultaneously regenerate functional alveoli. AEPs respond robustly to both Wnt and Fgf signals: Wnt signalling is a key factor in modulating the AT2-to-AT1 transition<sup>4</sup> and Fgfr2 activation promotes AT2 cell proliferation (Extended Data Fig. 10j). Importantly, the conservation and accessibility of both mouse and human AEPs provides an opportunity for

mechanistic studies to shed light on human lung progenitor cell biology, and assist in the development of new treatments for acute and chronic lung diseases.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 April 2017; accepted 24 January 2018.

Published online 28 February 2018.



**Figure 4 | AEPs display an evolutionarily conserved response to Wnt and Fgf signalling.** **a**, Human AEPs exhibit a distinct transcriptome enriched for Wnt responsiveness. **b**, More than one third of human AEP-enriched genes are shared with mouse AEPs. **c**, Volcano plot of 15,628 genes tested using limma shows extensive overlap between upregulated genes in mouse and human AEPs. FGFR2 is indicated. **d–w**, Alveolar organoid assays show that mouse AEPs (**d–m**) and human AEPs (**n–w**) display a significant increase in colony formation and size upon FGF7 treatment. Dep., AEP-depleted AT2 cells. Additional data are shown in Extended Data Fig. 10.  $n = 4$  individual organoid experiments. Statistics are inclusive of all biological replicates.  $**P < 0.01$  and  $***P < 0.001$  by ANOVA with adjustment for multiple comparison testing. NS, not significant. Plots are centred on mean with bars indicating standard deviation. Scale bars: 25  $\mu\text{m}$ .

- Beumer, J. & Clevers, H. Regulation and plasticity of intestinal stem cells during homeostasis and regeneration. *Development* **143**, 3639–3649 (2016).
- Stanger, B. Z. Probing hepatocyte heterogeneity. *Cell Res.* **25**, 1181–1182 (2015).
- Afelik, S. & Rovira, M. Pancreatic  $\beta$ -cell regeneration: facultative or dedicated progenitors? *Mol. Cell. Endocrinol.* **445**, 85–94 (2017).
- Frank, D. B. *et al.* Emergence of a wave of Wnt signaling that regulates lung alveologenesis by controlling epithelial self-renewal and differentiation. *Cell Reports* **17**, 2312–2325 (2016).
- Töpfer, L. *et al.* Influenza A (H1N1) vs non-H1N1 ARDS: analysis of clinical course. *J. Crit. Care* **29**, 340–346 (2014).
- Kumar, P. A. *et al.* Distal airway stem cells yield alveoli *in vitro* and during lung regeneration following H1N1 influenza infection. *Cell* **147**, 525–538 (2011).
- Vaughan, A. E. *et al.* Lineage-negative progenitors mobilize to regenerate lung epithelium after major injury. *Nature* **517**, 621–625 (2015).
- Zuo, W. *et al.*  $p63^+Krt5^+$  distal airway stem cells are essential for lung regeneration. *Nature* **517**, 616–620 (2015).
- Xi, Y. *et al.* Local lung hypoxia determines epithelial fate decisions during alveolar regeneration. *Nat. Cell Biol.* **19**, 904–914 (2017).
- Ray, S. *et al.* Rare SOX2<sup>+</sup> airway progenitor cells generate KRT5<sup>+</sup> cells that repopulate damaged alveolar parenchyma following influenza virus infection. *Stem Cell Reports* **7**, 817–825 (2016).
- Barkauskas, C. E. *et al.* Type 2 alveolar cells are stem cells in adult lung. *J. Clin. Invest.* **123**, 3025–3036 (2013).
- Rock, J. R. *et al.* Multiple stromal populations contribute to pulmonary fibrosis without evidence for epithelial to mesenchymal transition. *Proc. Natl Acad. Sci. USA* **108**, E1475–E1483 (2011).

- El-Hashash, A. H. *et al.* Six1 transcription factor is critical for coordination of epithelial, mesenchymal and vascular morphogenesis in the mammalian lung. *Dev. Biol.* **353**, 242–258 (2011).
- Herriges, J. C. *et al.* FGF-regulated ETV transcription factors control FGF–SHH feedback loop in lung branching. *Dev. Cell* **35**, 322–332 (2015).
- Kherrouche, Z. *et al.* PEA3 transcription factors are downstream effectors of Met signaling involved in migration and invasiveness of Met-addicted tumor cells. *Mol. Oncol.* **9**, 1852–1867 (2015).
- Rockich, B. E. *et al.* Sox9 plays multiple roles in the lung epithelium during branching morphogenesis. *Proc. Natl Acad. Sci. USA* **110**, E4456–E4464 (2013).
- Wan, H. *et al.* Kruppel-like factor 5 is required for perinatal lung morphogenesis and function. *Development* **135**, 2563–2572 (2008).
- Bogunovic, M. *et al.* Origin of the lamina propria dendritic cell network. *Immunity* **31**, 513–525 (2009).
- Wu, L. *et al.* MAT1-modulated CAK activity regulates cell cycle G<sub>1</sub> exit. *Mol. Cell. Biol.* **21**, 260–270 (2001).
- Lin, D. *et al.* Constitutive expression of B-myb can bypass p53-induced Waf1/Cip1-mediated G1 arrest. *Proc. Natl Acad. Sci. USA* **91**, 10079–10083 (1994).
- Schmidt-Edelkraut, U., Daniel, G., Hoffmann, A. & Spengler, D. Zac1 regulates cell cycle arrest in neuronal progenitors via Tcf4. *Mol. Cell. Biol.* **34**, 1020–1030 (2014).
- Marken, J. S., Schieven, G. L., Hellström, I., Hellström, K. E. & Aruffo, A. Cloning and expression of the tumor-associated antigen L6. *Proc. Natl Acad. Sci. USA* **89**, 3503–3507 (1992).
- Gao, H. *et al.* Multi-organ site metastatic reactivation mediated by non-canonical discoidin domain receptor 1 signaling. *Cell* **166**, 47–62 (2016).
- Gonzalez, R. F., Allen, L., Gonzales, L., Ballard, P. L. & Dobbs, L. G. HTII-280, a biomarker specific to the apical plasma membrane of human lung alveolar type II cells. *J. Histochem. Cytochem.* **58**, 891–901 (2010).
- Barkauskas, C. E. *et al.* Lung organoids: current uses and future promise. *Development* **144**, 986–997 (2017).
- Shu, W. *et al.* Wnt/ $\beta$ -catenin signaling acts upstream of N-myc, BMP4, and FGF signaling to regulate proximal-distal patterning in the lung. *Dev. Biol.* **283**, 226–239 (2005).
- Zhang, X. *et al.* Receptor specificity of the fibroblast growth factor family. The complete mammalian FGF family. *J. Biol. Chem.* **281**, 15694–15700 (2006).
- Yano, T. *et al.* KGF regulates pulmonary epithelial proliferation and surfactant protein gene expression in adult rat lung. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **279**, L1146–L1158 (2000).
- Yano, T., Deterding, R. R., Simonet, W. S., Shannon, J. M. & Mason, R. J. Keratinocyte growth factor reduces lung damage due to acid instillation in rats. *Am. J. Respir. Cell Mol. Biol.* **15**, 433–442 (1996).
- Panos, R. J., Rubin, J. S., Csaky, K. G., Aaronson, S. A. & Mason, R. J. Keratinocyte growth factor and hepatocyte growth factor/scatter factor are heparin-binding growth factors for alveolar type II cells in fibroblast-conditioned medium. *J. Clin. Invest.* **92**, 969–977 (1993).
- Quantius, J. *et al.* Influenza virus infects epithelial stem/progenitor cells of the distal lung: impact on Fgfr2b-driven epithelial repair. *PLoS Pathog.* **12**, e1005544 (2016).
- Nikolaidis, N. M. *et al.* Mitogenic stimulation accelerates influenza-induced mortality by increasing susceptibility of alveolar type II cells to infection. *Proc. Natl Acad. Sci. USA* **114**, E6613–E6622 (2017).
- Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported by grants from the National Institutes of Health (T32-HL007586 to W.J.Z.; T32-HL007915, K12-HD043245 to D.B.F.; T32-HL007843 to J.A.Z. and HL110942, HL087825, HL132999, HL129478, HL134745 to E.E.M.). We thank the Flow Cytometry Core Laboratory of Children's Hospital of Philadelphia and the CVI Histology Core, Next Generation Sequencing Core and CDB Microscopy Core at the University of Pennsylvania for technical assistance.

**Author Contributions** W.J.Z., D.B.F., J.A.Z., F.A.A., S.Z. and J.K. performed the experiments. W.J.Z., D.B.F., J.A.Z., M.P.M. and E.E.M. analysed the data. E.C. provided access to human samples and assisted W.J.Z. with all human experiments. E.E.M. supervised the project. W.J.Z. wrote the first draft of the manuscript. All authors contributed to the writing of the final manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to E.E.M. ([emorris@pennmedicine.upenn.edu](mailto:emorris@pennmedicine.upenn.edu)).

**Reviewer Information** Nature thanks C. Dean and the other anonymous reviewer(s) for their contribution to the peer review of this work.



## METHODS

**Ethical compliance.** All mouse studies were performed under guidance of the University of Pennsylvania Institutional Animal Care and Use Committee in accordance with institutional and regulatory guidelines. This study used cells derived from de-identified non-used lungs donated for organ transplantation via an established protocol (PROPEL, approved by University of Pennsylvania Institutional Review Board) with informed consent in accordance with institutional and NIH procedures. All patient information was removed before use. This use does not meet the current NIH definition of human subject research, but all institutional procedures required for human subject research were followed throughout the reported experiments.

**Mouse and Cre recombinase induction.** The generation and genotyping of the *Axin2<sup>creERT2-tdT</sup>* mouse line generated in our laboratory has previously been described<sup>4</sup>. The *Sftpc<sup>creERT2</sup>* mouse line was a gift of H. Chapman; their genotyping and generation have previously been described<sup>34</sup>. *Hoxp-3<sup>flagGFP</sup>* mice<sup>35</sup> were a gift of R. Jain and J. Epstein, and are available from Jackson Laboratories. The *R26R<sup>eYFP</sup>* mice are available from Jackson Laboratories. All mouse studies were performed under guidance of the University of Pennsylvania Institutional Animal Care and Use Committee. Mice were maintained on a mixed CD-1 and C57BL/6 background. For induction of all Cre recombinase models, tamoxifen (Sigma) was dissolved in 100% ethanol and diluted with corn oil (Sigma) to produce a 10% ethanol:tamoxifen:corn oil mixture at 20 mg/ml. Six-to-eight-week-old mice were injected intraperitoneally with 200 µg per g body weight on between three and five consecutive days to induce recombination. All lineage-tracing experiments represent a minimum of  $n = 6$  mice in all groups to allow for effective statistical evaluation. qPCR experiments represent a minimum of  $n = 3$  mice in all groups. Mouse experiments were performed on both male and female mice in all conditions, and mice were chosen at random from the cohort but not formally randomized. Blinding for experimental condition was not possible owing to the nature of the injury experiments.

**Influenza lung injury.** PR8 H1N1 influenza was a gift of J. Wherry. Recombination for lineage tracing was performed using 3 daily tamoxifen injections, 7 or 28 days before viral infection. For infection, the virus was diluted in PBS and a dose of 0.3 LD<sub>50</sub> was administered via intranasal instillation. After infection, mice were weighed and monitored daily for 14–28 days, and mice that lost >30% of their starting weight or were moribund were euthanized humanely. Post-influenza RNA was obtained at 14 days after infection, and lung regeneration was analysed from tissue collected from mice between 28 days and 3 months after infection. FACS data were generated from influenza-infected and uninfected mice using the same protocols, as described later. Regionalized lung injury was assessed via histology, and adjacent sections were used for all immunostaining and quantification.

**Histology.** At the time of tissue collection, mice were euthanized by CO<sub>2</sub> inhalation. The chest cavity was exposed and the lungs cleared of blood by perfusion with cold PBS via the right ventricle. Lungs were inflated with 2% paraformaldehyde under constant pressure of 30 cm water, and allowed to fix overnight. Tissue was then dehydrated, paraffin embedded and sectioned. Haematoxylin and eosin staining was performed to examine morphology, and to score regions on the basis of the severity of injury. Immunohistochemistry was used to detect protein expression, using the following antibodies on paraffin sections: GFP (chicken, Aves, GFP-1020, 1:500), GFP (goat, Abcam, ab5450, 1:100), RFP (rabbit, Rockland, 600-901-379, 1:250), Scgb1a1 (goat, Santa Cruz, sc-9772, 1:20), Tubb4 (mouse, BioGenex, MU178-UC, 1:20), Sftpc (rabbit, Millipore, ABC99, 1:250), Sftpc (goat, Santa Cruz, sc-7750, 1:50), Pdpn (mouse, HybriDoma Bank, Clone 8.1.1, 1:50), Aqp5 (rabbit, Abcam, ab92320, 1:100), Ki67 (rabbit, Abcam, clone SP6, ab16667, 1:50) and anti-mouse Tm4sf1 (rabbit, LSBiosciences, B7077, 1:500).

**Alveolar epithelial cell number and lineage imaging and quantification.** After immunostaining for alveolar epithelial lineages and proliferation, images were captured using a Nikon Eclipse Ni wide-field microscope or a Leica TCS SP8 confocal microscope. We captured images containing at least eight individual 1-µm optical sections. Z-stacks were obtained from at least five random areas of each histological zone from a minimum of  $n = 5$  mice. All images were processed with ImageJ software. Cell counts were performed using the Cell Counter plug-in for ImageJ. Cells were counted in at least three different areas of each histological injury zone for each mouse, to obtain a total count of >1,000 cells counted for each condition. Only true confocal images were used for quantification. For image presentation, both confocal images and images obtained with automatic deconvolution algorithms in Nikon Elements software are presented, with source as noted.

**Lung alveolar epithelial cell isolation and FACS analysis.** *Mouse.* Lungs from *Axin2<sup>creERT2-tdT</sup>* mice were collected at 6–8 weeks of age and processed into a single-cell suspension using dispase, collagenase I and DNase, as previously described<sup>36,37</sup>. EPCAM<sup>+</sup>Axin2<sup>+</sup> cells (tdTomato<sup>+</sup>) were identified via FACS sorting, as previously described<sup>37</sup>. The total AT2 population (Sftpc<sup>+</sup> AT2 cells) was isolated from lungs of 6–8-week-old *Sftpc<sup>creERT2</sup>;R26R<sup>eYFP</sup>* mice 5 days after

induction with 200 µg per g body weight tamoxifen. eYFP<sup>+</sup> cells were then isolated via FACS sorting, as previously described<sup>37</sup>. For sorting and quantification, the following antibodies were used: Pdpn-eFluor660 (eBioscience, Clone 8.1.1, 1:100), EpCAM-APC (eBioscience, Clone G8.8, 1:200), EpCAM-eFluor488 (eBioscience, Clone G8.8, 1:200), CD31-PeCy7 (eBioscience, Clone 390, 1:200) and CD45-PeCy7 (eBioscience, Clone 30-F11, 1:200). Two anti-mouse Tm4sf1 antibodies were used to ensure specificity: sheep anti-mouse Tm4sf1 (R&D systems, AF7514, 1:10) and sheep IgG isotype control (R&D systems, 5-001-A, 1:10) with anti-sheep 488 secondary (Abcam, ab150177, 1:50) or rabbit anti-mouse Tm4sf1 (LS Biosciences, B7077, 1:25) and rabbit IgG isotype control (LS Biosciences, LS-C109221, 1:25) with donkey anti-rabbit 488 secondary (Life Technologies, A212016, 1:200).

*Human.* Samples of normal, de-identified human lungs were obtained from non-used lungs donated for organ transplantation via an established protocol (PROPEL, approved by University of Pennsylvania Institutional Review Board) with informed consent in accordance with institutional procedures. A 2 × 2 cm piece of distal lung tissue was obtained, pleura and large airways were carefully dissected away and tissue was processed into a single-cell suspension using the same combination of dispase, collagenase I and DNase that was used for mouse lungs. A Miltenyi gentleMACS dissociator was used for mincing and incubation for 35 min at 37°C. Cells were washed, passed over 70-µm and 40-µm filters and red blood cells were lysed with ACK lysis buffer. After a single-cell suspension was obtained, cells were analysed by FACS or sorted using the MACS multisort kit, MACs LS columns and the following antibodies: EPCAM-PE (BD, mouse, Clone 1B7, 1:50), HT2-280 (mouse IgM, a gift of L. Dobbs, UCSF, 1:50), TM4SF1-APC (mouse, R&D Systems, Clone 877621, 1:100), Mouse IgG1-APC isotype control (R&D systems, 1C002A, 1:100), anti-APC microbeads (Miltenyi, 130-090-855, 1:20) and anti-mouse IgM microbeads (Miltenyi, 130-047-302, 1:20). For the full protocol for digestion and sorting of human lung epithelial cells, and their propagation as alveolar organoids, see 'Data availability'.

**Ex vivo alveolar organoids.** Clonal alveolar organoid assays were performed as previously described, with some modifications from the original protocol<sup>4,11,25,37</sup>. In brief, 5 × 10<sup>3</sup> epithelial cells (AT2 or AEP for mouse, HT2-280<sup>+</sup>, HT2-280<sup>+</sup>TM4SF1<sup>+</sup> and HT2-280<sup>+</sup>TM4SF1<sup>-</sup> for human) were isolated as described earlier, and mixed with 5 × 10<sup>4</sup> lung fibroblasts (isolated from adult wild-type mice as previously described<sup>37</sup> for mouse; MRC5 cells (ATCC CCL-171, tested negative for mycobacterial contamination, at no greater than passage 10) for human). Cells were then suspended in a 1:1 mixture of SAGM medium (Lonza, with all additives except epinephrine) and growth factor-reduced, phenol-free Matrigel (Corning). Ninety microlitres of the cell–medium–Matrigel mixture was then aliquoted into individual 24-well cell culture inserts and allowed to solidify at 37°C. SAGM was then placed into each well of the 24-well plate. The Rock inhibitor Y27632 (Sigma) was included in the medium for the first two days. After two days of culture, Y27632 was removed and ligand treatments of organoids were performed using the following reagents at the indicated concentrations: Wnt3a 200 ng/ml (R&D systems), Fgf7 50 ng/ml (R&D Systems), Fgf10 50 ng/ml (R&D Systems), XAV939 10 µM (Sigma) and CHIR99021 1 µM (Fisher). DMSO was used as a control. The medium was changed every 48 h, and fresh ligands were included at each medium change. After 21 days of culture, organoids were fixed in 2% paraformaldehyde, embedded in Histogel (Richard-Allan), dehydrated, paraffin-embedded, and sectioned and immunostained as described earlier.

**RNA-seq analysis.** Cells were sorted using the protocols described earlier into Trizol LS (Life Technologies). For mouse, six individual mice were sorted and pooled into two individual pools for *Axin2<sup>+</sup>* cells and three individual pools for *Sftpc<sup>+</sup>* cells. For human, cells from three individual patients were sorted separately and prepared for sequencing individually. RNA was then extracted using a combination of the Trizol protocol and MinElute RNA Cleanup Kit (Qiagen). RNA integrity was confirmed via Bionalyzer evaluation and samples with RNA integrity numbers >8.5 were chosen for library preparation. Library preparation was conducted using Illumina truSeq stranded mRNA kit, followed by the Nugen Ovation amplification kit. FASTQ files were assessed for quality control using the FastQC program. FASTQ files were aligned against the mouse reference genome (mm9) or human reference (hg19/hGRC37) genome using the STAR aligner<sup>38</sup>. Duplicate reads were flagged using the MarkDuplicates program from Picard tools. Per-gene read counts for Ensembl (v.67) gene annotations for the mouse samples or Ensembl (v.75) for the human samples were computed using the R package Rsubread, with duplicate reads removed. Gene counts represented as counts per million (CPM) were first normalized using the trimmed mean of *M*-values method in the R package edgeR, and genes with a CPM < 1 in 25% of samples were removed and deemed to be low-expressed. These data were transformed using the VOOOM function from the limma R package<sup>39</sup>. Differential gene expression was performed using a linear model with the limma package. Given the small sample size of the experiment, we employed the empirical Bayes procedure as implemented in limma to adjust the linear fit and calculate *P* values. *P* values were adjusted for



multiple comparisons using Benjamini–Hochberg procedure. For the human data, a paired analysis was employed using the patient as a blocking variable. Heat maps and principal component analysis plots were generated in R. Gene Ontology (GO) enrichment analysis was performed using the ToppGene Suite (<http://toppgene.cchmc.org/>)<sup>40</sup>.

**ATAC-seq analysis.** Individual ATAC-seq libraries were generated from sorted Axin2<sup>+</sup> and Sftpc<sup>+</sup> AT2 cells as described earlier, using previously published methods<sup>41</sup>. In brief,  $5 \times 10^4$  cells were sorted into PBS, washed and lysed to obtain nuclei. Nuclei were exposed to Tn5 transposase (Illumina), and fractionated DNA was used for amplification and library preparation. Libraries were then purified and paired-end sequenced. After sequencing, FASTQ files were aligned against the mouse reference genome (mm9) using the STAR aligner<sup>38</sup>, with default parameters plus options to suppress the matching of spliced reads ('-outFilterMatchNminOverLread 0.4--outFilterScoreMinOverLread 0.4'). Duplicate reads were flagged using the MarkDuplicates program from Picard tools and removed using samtools. MACS2<sup>42</sup> was used to call peaks with the following options '-nomodel-shift -100-extsize 200'. Differential ATAC-seq peaks were determined using the bddiff command from MACS2 and default options. Peaks were filtered to have a MACS2 log<sub>10</sub> likelihood ratio score >10 and to be within -50 kb and 10 kb of the transcription start site of Ensembl 67 protein-coding genes. ATAC-seq enrichment heat maps were created using deepTools<sup>43</sup>.

**Motif analysis.** The intersection of gene promoter regions (-5 kb, 600 bp, Ensembl v.67) with identified ATAC-seq peaks was performed using bedtools. The FASTA file of genome sequence (mm9) of promoter ATAC-seq peaks was created using bedtools and scanned for TCF/LEF motifs using FIMO<sup>44</sup>. Motif enrichment analysis was performed using the findMotifsGenome.pl program in the HOMER software suite<sup>45</sup>, with the peak search size option set to 50 bp.

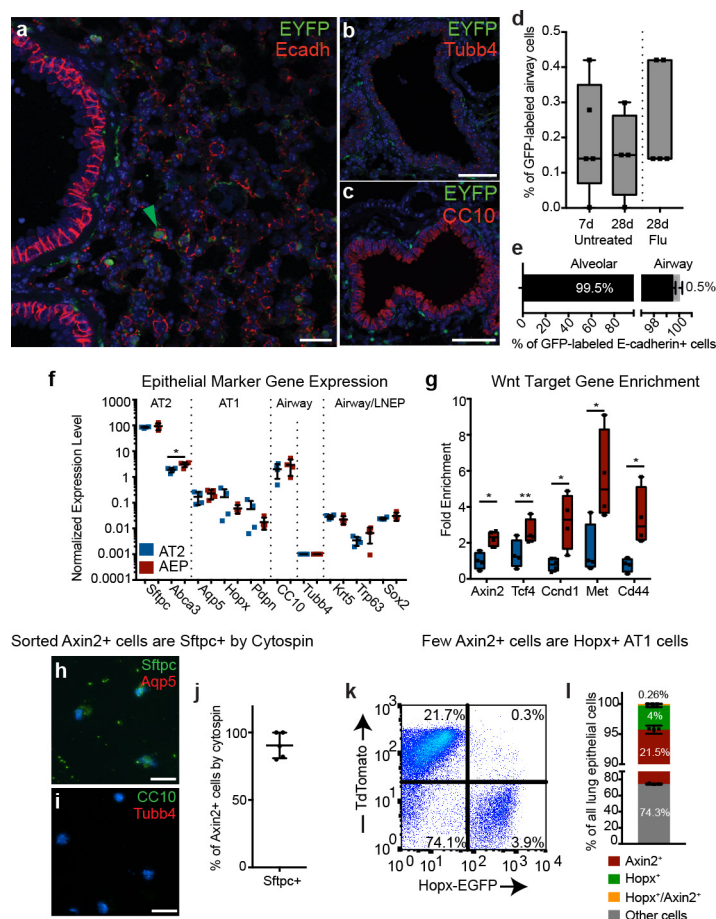
**Chromatin immunoprecipitation.** Chromatin immunoprecipitation was performed using the High Sensitivity ChIP Kit (Abcam) with 3 µg of anti-β-catenin (Santa Cruz sc-7963) or anti-IgG1 isotype control (Santa Cruz sc-3877). In brief,  $1 \times 10^5$  Axin2<sup>+</sup> or Sftpc<sup>+</sup> AT2 cells were sorted into SAGM (Lonza), whole chromatin was prepared, chromatin was cross-linked and sonicated using a Covaris sonicator to an optimal size of 300 bp, and chromatin was immunoprecipitated using the antibodies listed earlier, following the Abcam protocol. Library quality was confirmed using Bioanalyzer, and the enrichment of genomic DNA was assessed using qPCR to compare β-catenin immunoprecipitate to immunoprecipitate from an IgG control for each cell type. qPCR data represents  $n = 2$  individual immunoprecipitation experiments, and was performed in triplicate.

**Statistical analysis.** No statistical methods were used to predetermine sample size. Statistical analysis was performed in Prism for Mac, and R. A two-tailed Student's *t*-test was used for the comparison between two experimental groups. For experiments with more than two groups, an ANOVA was performed followed

by planned contrasts; pairwise comparisons and *P*-value adjustments for multiple comparisons were performed using Dunnett's procedure. The generation of odds ratios for the distribution of ATAC regions near genes was evaluated using Fisher's exact test and contingency table analysis. Statistical data were considered significant when  $P < 0.05$ . Centre values of all plots represent means and error bars represent standard deviations, with the exception of error bars for odds ratios (which represent confidence intervals).

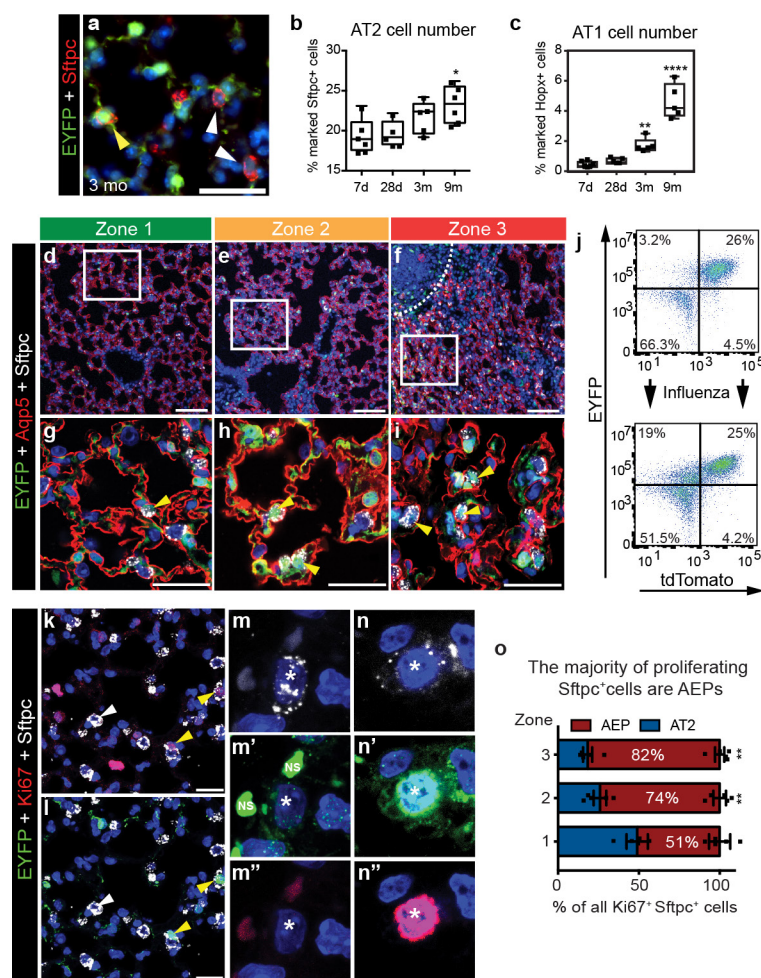
**Data availability.** ATAC-seq and RNA-seq sequencing data generated during this study have been deposited in the Gene Expression Omnibus database with the primary accession GSE97055. All upregulated and downregulated genes identified during the RNA-seq experiments described in this paper are found in Supplementary Table 1 (mouse data) or Supplementary Table 2 (human data). Source Data for all plots in all figures are available online. The detailed protocol for the cell isolation and propagation of human AEPs is available on the Protocol Exchange<sup>36</sup>. All other datasets generated during and/or analysed in the current study are available from the corresponding author on request.

34. Chapman, H. A. *et al.* Integrin α6β4 identifies an adult distal lung epithelial population with regenerative potential in mice. *J. Clin. Invest.* **121**, 2855–2862 (2011).
35. Takeda, N. *et al.* Hopx expression defines a subset of multipotent hair follicle stem cells and a progenitor population primed to give rise to K6<sup>+</sup> niche cells. *Development* **140**, 1655–1664 (2013).
36. Zacharias, W. J. & Morrissey, E. E. Isolation and culture of human alveolar epithelial progenitor cells. *Protoc. Exch.* <http://dx.doi.org/10.1038/protex.2018.015> (2018).
37. Peng, T. *et al.* Hedgehog actively maintains adult lung quiescence and regulates repair and regeneration. *Nature* **526**, 578–582 (2015).
38. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
39. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
40. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).
41. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.21–21.29.9. (2015).
42. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
43. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
44. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
45. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).



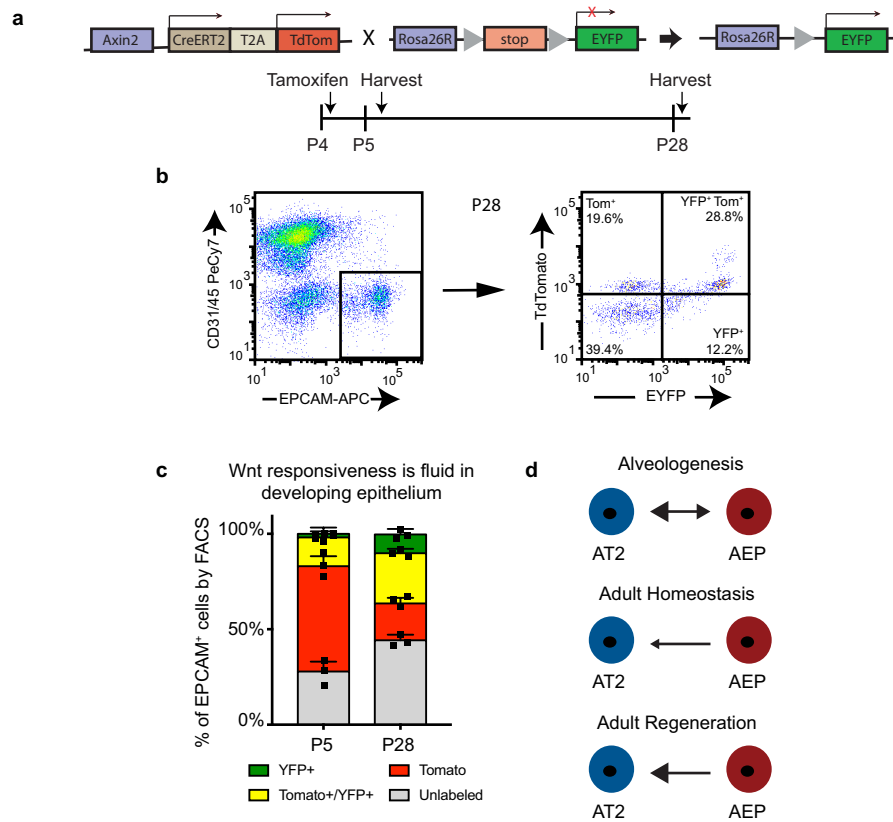
**Extended Data Figure 1 | Location of Axin2<sup>+</sup> epithelial cells within the adult mouse lung.** **a**, Low-power view of the lung showing that E-cadherin<sup>+</sup>Axin2<sup>+</sup> epithelial cells are found only in the alveolar region, and not in the airway of the lung. **b, c**, Immunohistochemistry for ciliated (**b**) and secretory (**c**) markers shows no evidence of Axin2-lineage labelled cells co-expressing either of these markers. **d, e**, Quantification of the location of Axin2<sup>+</sup> epithelial cell distribution in the lung. **f**, qPCR showing that Axin2<sup>+</sup> AEPs and AT2 cells express similar levels of AT2 markers and other lung epithelial cell markers. AEPs express slightly higher levels of *Abca3*. **g**, AEPs express increased levels of Wnt signalling pathway components and targets by qPCR. **h-j**, Cytopsin and

quantification demonstrating that the majority of sorted Axin2<sup>+</sup> epithelial cells are Sftpc<sup>+</sup>. **k, l**, FACS analysis of Axin2<sup>tdT</sup>-positive, *Hopx*<sup>eYFP</sup> mice demonstrating that few Axin2<sup>+</sup> epithelial cells express Hopx, consistent with the immunohistochemistry data shown in Fig. 1. Data in this figure represent  $n = 3$  (**k, l**), 4 (**d-j**) or 10 (all other panels) mice from three individual experiments. Statistics are representative of all biological replicates. All data are shown as centred on mean with bars indicating standard deviation. \* $P < 0.05$ , \*\* $P < 0.01$  by two-tailed  $t$ -test (**f, g**) or ANOVA with preplanned pairwise comparisons and adjustment for multiple comparison testing (**d**). Scale bars: **a-c**, 100  $\mu$ m; **h, i**, 25  $\mu$ m.



**Extended Data Figure 2 | Characterization of Axin2<sup>+</sup> Wnt responsive cells in the adult lung.** **a**, Lineage tracing for three months shows a stable population of AEPs and progeny in the alveolar epithelium. Yellow arrow, labelled cell; white arrow, unlabelled cell. **b**, **c**, Quantification of AT1 and AT2 cells labelled by the AEP lineage mark at homeostasis. Lower power (**d–f**) and higher power (**g–i**) images showing expansion of AEPs in a regional fashion, one month after influenza injury. Dotted white line in **f** shows the edge of a Krt5<sup>+</sup> pod, with a dearth of AEP-lineage-labelled cells. Panels **g–i** show additional channels of the same fields as shown in Fig. 1i, j. **j**, Representative FACS plot showing expansion of AEP-lineage-labelled epithelial cells after influenza. The quantification of these FACS plots can

be found in Fig. 1n. **k–o**, Comparison of Ki67<sup>+</sup> expression in AT2 cells and AEPs after influenza. In areas of regeneration, Ki67<sup>+</sup> AEPs constitute the majority of cells entering the cell cycle, when compared to AT2 cells. Data shown represent  $n = 5$  (**j–o**), 6 (**a–c**) or 10 (**d–i**) independent mice from three individual experiments, except for the nine-month lineage tracing which was performed in two separate experiments. Statistics are representative of all biological replicates. All data are shown as centred on mean with bars indicating standard deviation. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  and \*\*\*\* $P < 0.0001$  by ANOVA with preplanned pairwise comparisons and adjustment for multiple comparison testing. Scale bar, 50  $\mu\text{m}$ .

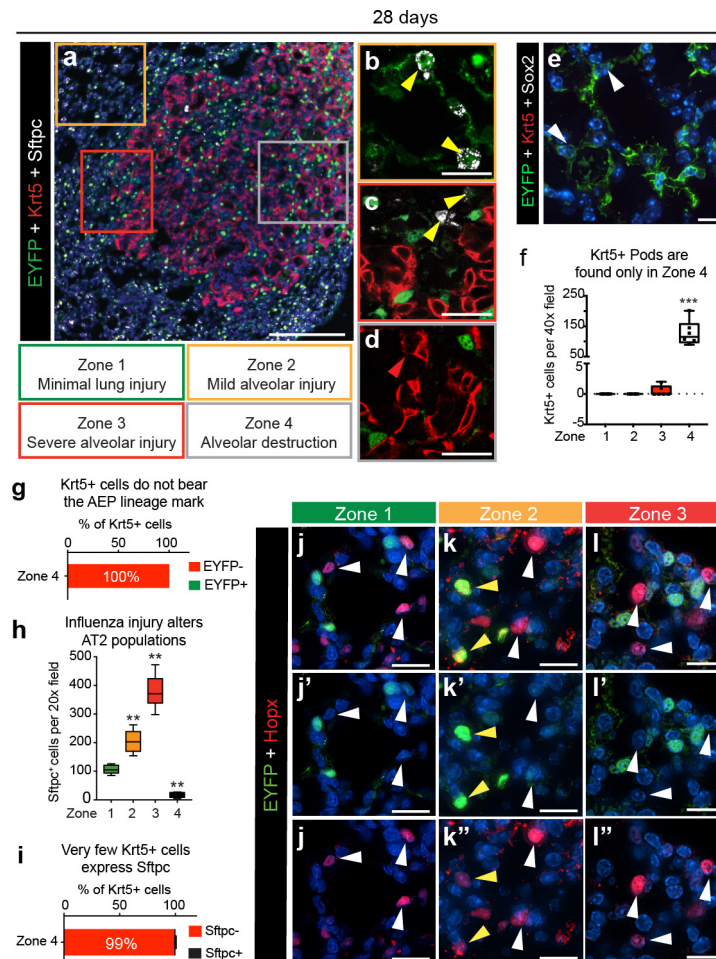


**Extended Data Figure 3 | In contrast to adult lung homeostasis, the Wnt response in the alveolar epithelium during alveologenesis is dynamic.**

**a**, Schematic of lineage labelling procedure to assess Wnt-responsive epithelium during alveologenesis. **b**, Epithelial cells were identified by FACS as  $Epcam^{+}CD45^{-}CD31^{-}$ . Cells were then gated for tdTomato and eYFP expression as shown. **c**, Quantification of Wnt responsiveness in the alveolar epithelium over a 1-day or 3-week lineage trace. **d**, Model of directionality and magnitude of AT2 and AEP transitions. During

alveologenesis, AT2 and AEP fates are somewhat fluid, though the AEP population decreases during this period of lung development. During adult homeostasis, few if any AT2 cells take on the AEP fate (see Fig. 2). After injury, AEPs expand to create AT2 cells, but even after injury very few AT2 cells adopt the AEP fate. Data shown represent  $n = 3$  mice. Statistics are representative of all biological replicates. Data in **c** are centred on mean with bars indicating standard error of the mean.

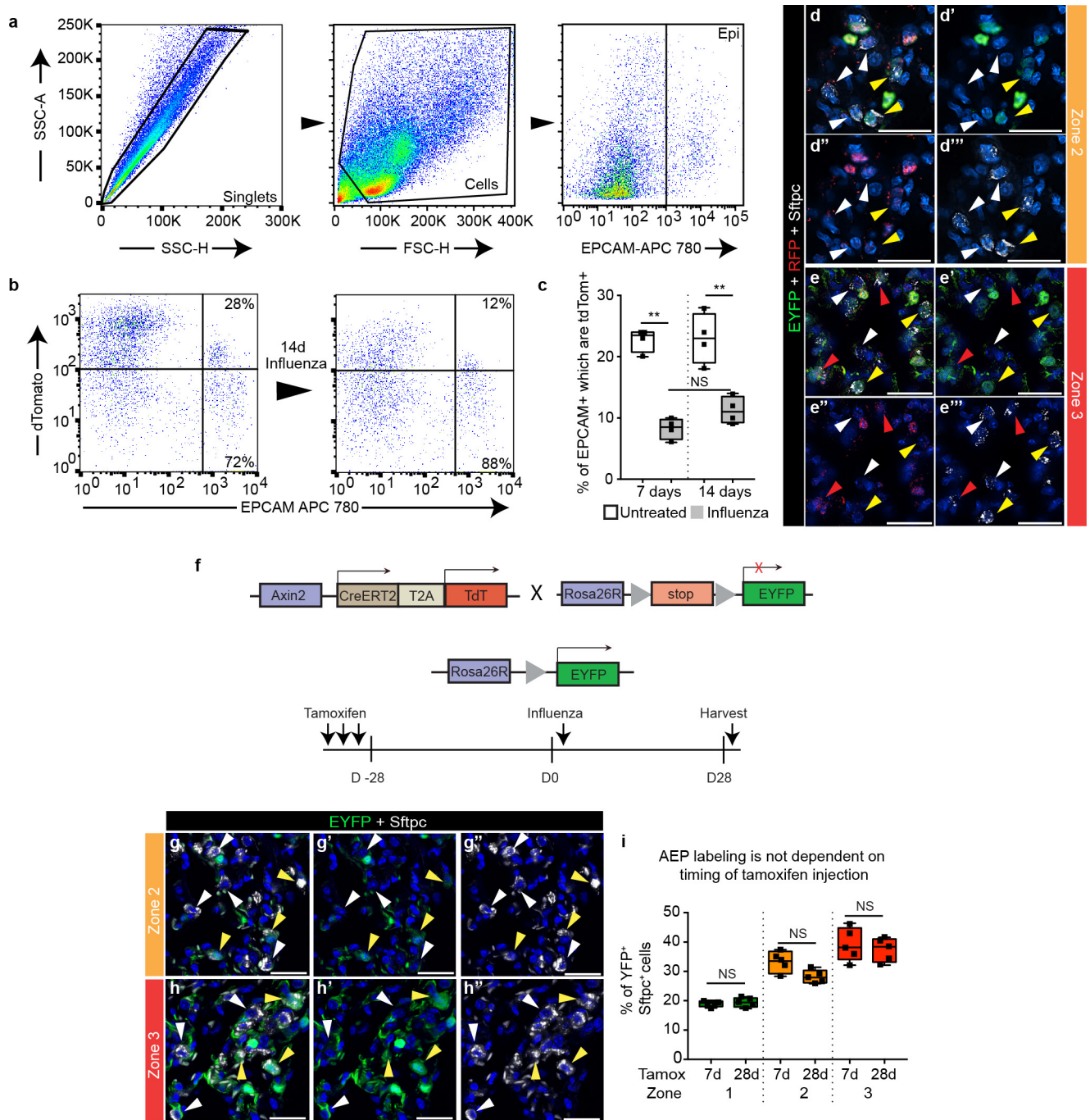




**Extended Data Figure 4 | AEPs are a distinct lineage compared to Sox2-derived Krt5<sup>+</sup> cells and are capable of generating AT1 cells.**

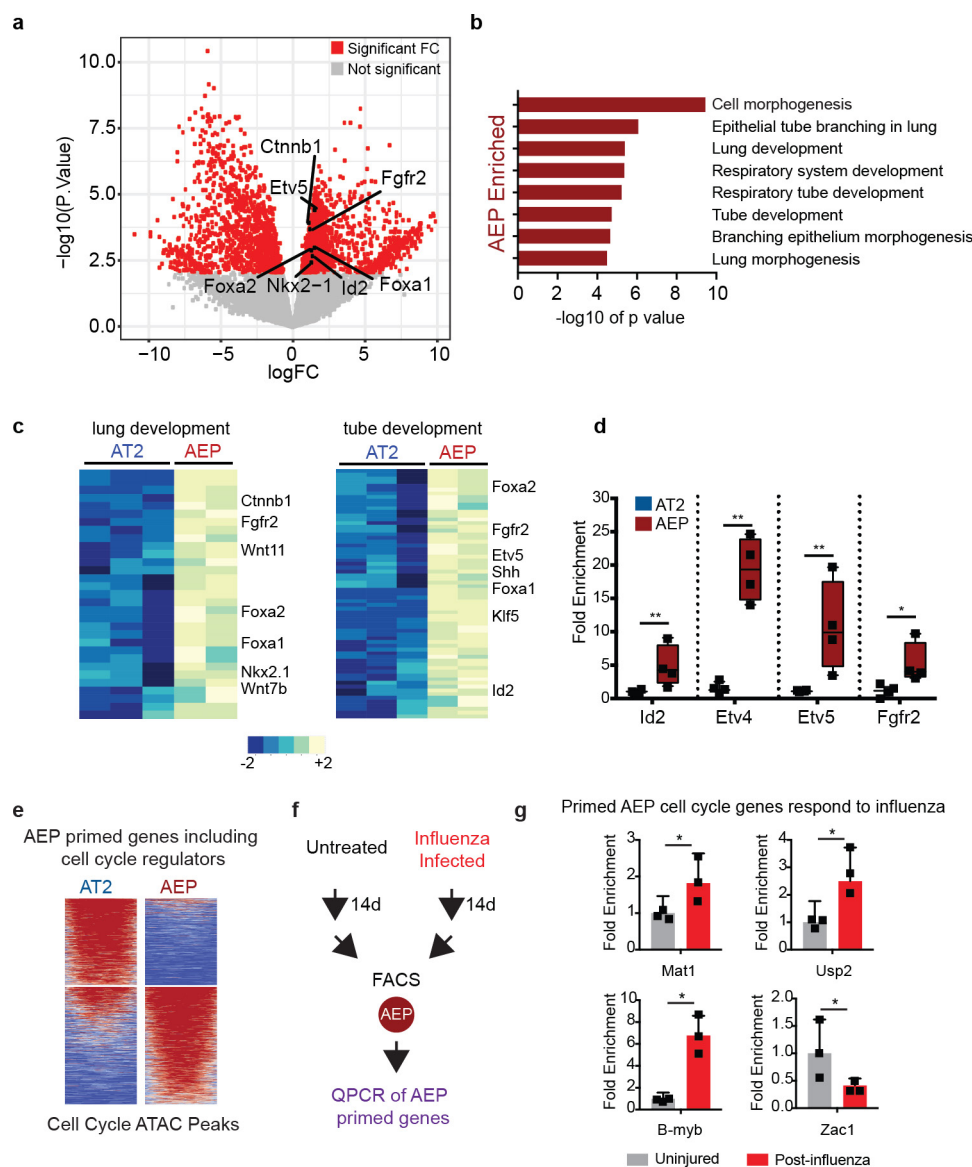
**a–d**, AEPs and Krt5<sup>+</sup> cells inhabit distinct regions of the regenerating mouse lung. **a**, Overview of a region surrounding a Krt5<sup>+</sup> pod. **b**, In regions of mild injury, AEPs and AEP-lineage-marked AT2 cells predominate and no Krt5<sup>+</sup> cells are seen. Yellow arrow, AEP-labelled cell. **c**, At the border of zone 4 areas of alveolar destruction, AEPs are observed regenerating AT2 cells. **d**, Krt5<sup>+</sup> cells are distinct from AEPs and never bear the AEP lineage mark. Red arrow, Krt5<sup>+</sup> cell. **e**, AEP-lineage cells do not express Krt5 or Sox2 protein at baseline, in contrast to previously reported lineages<sup>7,8</sup>. Arrows represent probable AEPs by morphology. **f**, Krt5<sup>+</sup> cells predominate in zone 4 regions, where AEPs are not present. **g**, Quantification demonstrating that Krt5<sup>+</sup> cells are never marked

with the AEP lineage mark. **h**, AT2 populations expand markedly after influenza injury, except in zone 4. **i**, Krt5<sup>+</sup> cells rarely express Sftpc in zone 4 regions. **j–l**, One month after influenza injury, AEPs give rise to a small number of Hopx<sup>+</sup> AT1 cells, predominantly in zone 2 of mild injury. Yellow arrow, AEP-labelled cells; white arrow, unlabelled cells. Zone 3 (l) has very few AEP-derived Hopx<sup>+</sup> cells, which may be due to a lag in AT1 regeneration from AEPs in this more severely affected region. Data shown represent  $n = 6$  (a–g, i) or 10 (h, j, k) independent mice across three individual experiments. Statistics are representative of all biological replicates. All data are shown as centred on mean with bars indicating standard deviation.  $^{**}P < 0.01$  and  $^{***}P < 0.001$ , by ANOVA with preplanned pairwise comparisons and adjustment for multiple comparison testing. Scale bars: a, 200  $\mu\text{m}$ ; b–d, j–l, 50  $\mu\text{m}$ .



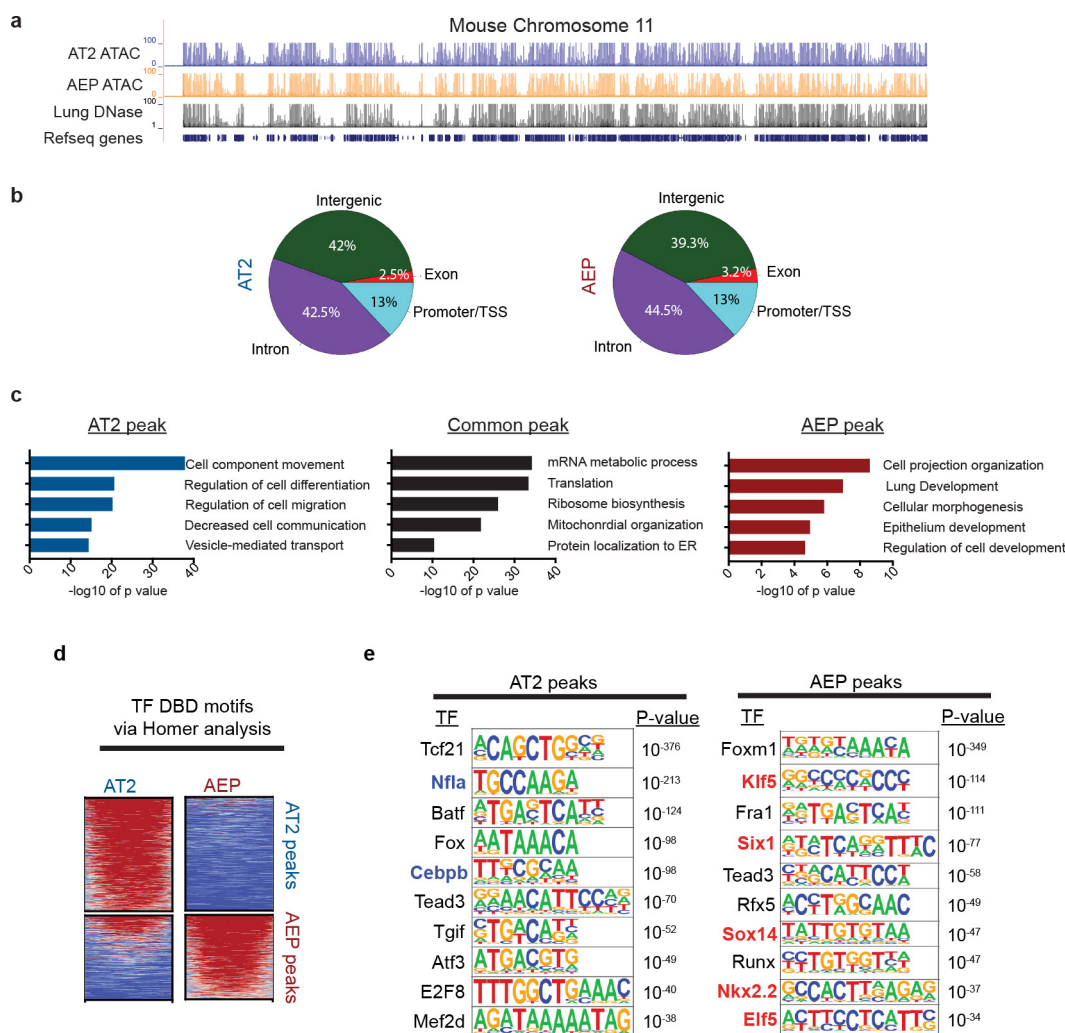
**Extended Data Figure 5 | Wnt signalling in the alveolar epithelium is largely stable after influenza infection, and AEP lineage labelling is not affected by tamoxifen perdurance.** **a**, FACS gating strategy used for all post-influenza FACS experiments in Fig. 1, Extended Data Fig. 2 and **b**, **c**. SSC-A, side-scatter area, SSC-H, side-scatter height, FSC-H, forward-scatter height. **b**, **c**. FACS analysis demonstrates that  $\text{Axin2}^{\text{tdT}}$  intensity is mildly decreased in the epithelium at 7 and 14 days after influenza infection. **d**, In regions of milder lung injury, most lineage-labelled AT2 cells are  $\text{eYFP}^+$  and  $\text{tdTomato}^-$ , which suggests that these cells are the progeny of AEPs. **e**, In zone 3, we detect a mix of  $\text{eYFP}^+\text{tdTomato}^+$  AEPs (red arrowheads) and  $\text{eYFP}^+\text{tdTomato}^-$  AEP progeny (yellow arrowheads) among the AT2 cell population. **f**, Experimental design of lineage tracing experiment in **g**–**i**, with a longer incubation time after tamoxifen treatment

than in the experiments that generated the data presented in **a**–**e**, and Fig. 1 and Extended Data Figs 4, 6. **g**, **h**, Confocal imaging demonstrating lineage labelling of AT2 cells with the AEP lineage mark 28 days after influenza-mediated injury. White arrows, unlabelled AT2 cells; yellow arrows, AEP-labelled cells. **i**, Quantification of lineage-labelled AT2 cells in multiple regions of lung injury. Representative seven-day lineage data is reproduced from Fig. 1 for comparison. Data shown represent  $n = 4$  (**a**–**c**) or 5 (**d**–**i**) independent mice across two different experiments. Statistics are representative of all biological replicates. All data were analysed with ANOVA followed by preplanned pairwise comparisons and adjustment for multiple comparison testing, and are shown centred on mean with bars indicating standard deviation. \*\* $P < 0.01$ . Scale bars, 50  $\mu\text{m}$ .



**Extended Data Figure 6 | Transcriptome analysis of AEPs versus AT2 cells, and activation of cell-cycle genes in AEPs after influenza injury.** **a**, Volcano plot of 14,618 genes tested using a linear model in the R package limma, showing the distinct differences in gene expression in AEPs and AT2 cells. Notable lung-progenitor developmental signalling and transcription factors are indicated. **b**, GO analysis of the top 500 most-differentially expressed genes, showing the enrichment of categories related to lung development and morphogenesis in AEPs. **c**, Heat maps of two of the AEP-enriched GO categories. Important regulators of lung-progenitor-cell biology are indicated. **d**, qPCR confirms upregulation of a

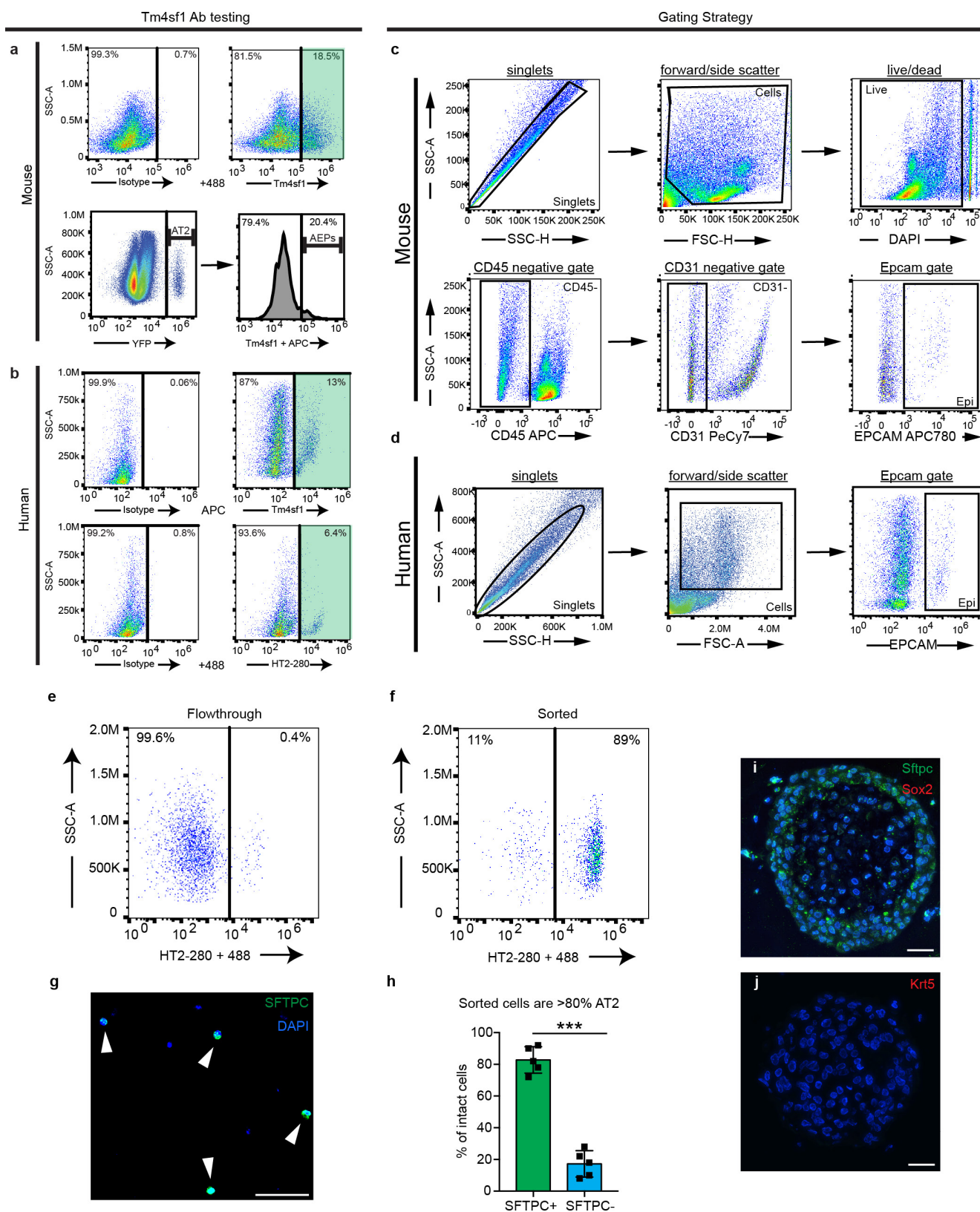
subset of important regulators of lung progenitor biology in AEPs. **e**, AT2 and AEP open chromatin is found near distinct sets of genes involved in the cell cycle. **f**, Schematic of analysis of changes in expression of AEP-primed genes after influenza infection. **g**, A subset of primed cell-cycle regulators in AEPs show expression changes after influenza infection. qPCR data are from  $n = 4$  mice from two separate infections. All data are shown as centred on mean with bars indicating standard deviation. Statistics are representative of all biological replicates. \* $P < 0.05$  and \*\* $P < 0.01$  by two-tailed  $t$ -test.



**Extended Data Figure 7 | ATAC-seq reveals distinct differences in open chromatin architecture in AEPs versus AT2 cells.** **a**, ATAC-seq peaks in both AT2 cells and AEPs are similar to previously described<sup>34</sup> mouse lung genome-wide DNase hypersensitivity profiling. **b**, AT2 and AEP ATAC peaks are distributed in a similar fashion, predominantly within intergenic regions and introns. **c**, GO enrichment analysis of the nearest neighbour genes in the vicinity of AT2 peaks, AEP peaks and peaks common to both AEPs and AT2 cells shows that common peaks are enriched for general cellular housekeeping roles, whereas AT2 open chromatin is enriched near

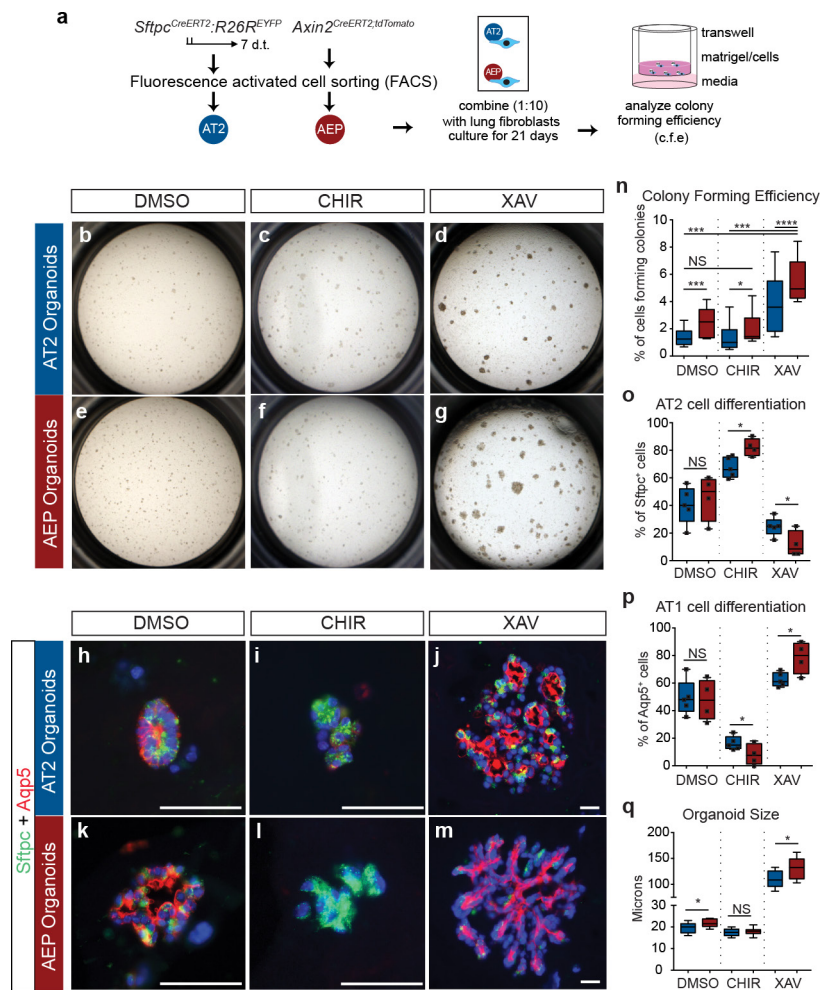
genes associated with exocytosis and cell differentiation. By contrast, AEP peaks are enriched near genes associated with lung development processes. **d**, **e**, Examination of the genes associated with open chromatin in AEPs reveals a strong enrichment for transcription factors associated with lung endoderm progenitor cells, including members of Klf, Six, Sox, Nkx2 and Elf/Ets families. By contrast, AT2 cell open chromatin is associated with a unique set of transcriptional regulators that includes members of the Nfl and Cebp families, which are known to regulate AT2 cell surfactant genes. For details of ATAC analysis, see Methods.





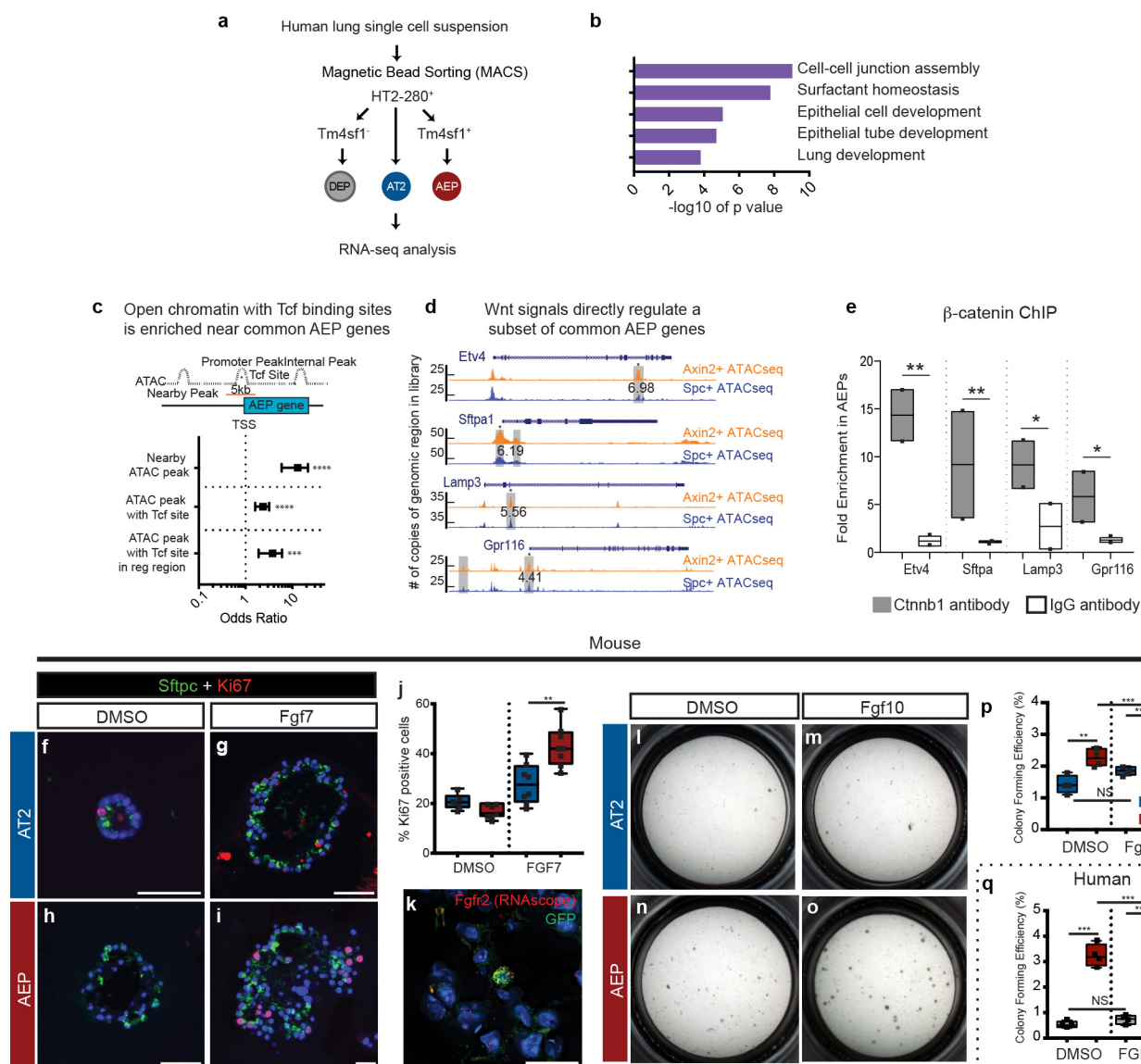
**Extended Data Figure 8 | The combination of HT2-280 and TM4SF1 antibodies are capable of identifying AEPs in human lung. a**, Top panels show isotype and active antibody gates for sheep anti-mouse Tm4sf1 FACS. The bottom panels show that the Tm4sf1 antibody detects approximately 20% of *Sftpc*<sup>creERT2eYFP</sup> labelled AT2 cells. **b**, Isotype and active antibody gates for human HT2-280 (AT2 marker) antibody and TM4SF1 antibody. **c**, **d**, An example of the FACS gating strategy used to generate the data shown in Fig. 3. **e**, **f**, Selection for HT2-280 strongly enriches for human AT2 cells. **g**, **h**, The majority of isolated HT2-280

cells express SFTPC protein by cytospin. **i**, **j**, Human AEPs in organoid culture do not express KRT5 or SOX2 protein at detectable levels. Each FACS panel shown in **a**–**f** shows gates from cells of one individual mouse or patient and is representative of  $n=6$  independent mice across two individual experiments or  $n=4$  human patients. Isotype staining was performed three times to confirm specificity. Statistics are representative of all biological replicates. Statistics in **h** are calculated with two-tailed *t*-test, displayed as mean with bars showing standard deviation. Scale bars: **g**, 25  $\mu$ m; **i**, **j**, 50  $\mu$ m (**i**, **j**).



**Extended Data Figure 9 | Mouse AEPs generate more alveolar organoids compared to AT2 cells, and cells in these organoids are restricted from AT1 cell differentiation by Wnt signalling.** **a**, Schematic of mouse alveolar organoid culture method. **b–m**, Sftpc<sup>+</sup> mouse AT2 cells (**b–d**, **h–j**) and mouse AEPs (**e–g**, **k–m**) were isolated from the indicated mouse lines and cultured in alveolar organoid assays. AT2 cells (**b**) and AEPs (**e**) both form alveolar organoids. AEPs generate more numerous and larger organoids than do AT2 cells. Activation of Wnt signalling using CHIR99021 does not increase the organoid-forming efficiency of either AT2 cells (**c**) or AEPs (**f**) but does increase the number of Sftpc<sup>+</sup> cells in treated organoids (**i**, **l**, **o**). Inhibition of Wnt signalling using XAV939 increases the number and size of alveolar organoids (**d**, **g**, **n**, **q**), decreases

the number of Sftpc<sup>+</sup> AT2 cells and increases the number of Aqp5<sup>+</sup> AT1 cells (**j**, **m**, **p**). For tests of all parameters, AEPs exhibited a more marked response to Wnt modulation than did AT2 cells. Data shown represent  $n = 12$  wells from  $n = 4$  individual mice in each group, across 3 individual experiments. Quantitative counting shown for cell differentiation (**o**, **p**) represents counting of  $n > 400$  organoids from  $n = 4$  mice. All data were analysed with ANOVA followed by preplanned pairwise comparisons and adjustment for multiple comparison testing, and are shown centred on mean with bars indicating standard deviation. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  and \*\*\*\* $P < 0.0001$ . Statistics are representative of all biological replicates. Scale bars: 50  $\mu\text{m}$ .



**Extended Data Figure 10 | Combination of ATAC-seq and RNA-seq emphasizes the Wnt- and FGF-responsive nature of AEPs and identifies several novel AEP-enriched direct Wnt target genes.** **a**, Schematic of human RNA-seq experiments. **b**, GO term analysis of the top 300 human AEP-enriched genes shows enrichment of several categories associated with lung progenitor cell function, similar to observations made of mouse AEPs. **c**, Evaluation of chromatin accessibility in the mouse genome near common AEP-enriched genes demonstrates a significant overrepresentation of Tcf binding sites, particularly in putative regulatory regions 5 kb immediately upstream of the transcriptional start site. For details of enrichment analysis, see Methods. **d**, Schematic of areas of AEP-enriched open chromatin near selected AEP-enriched genes. Peak height represents coverage of the indicated genomic region in the ATAC library, and the number indicates the fold enrichment in the indicated peak. **e**, Chromatin immunoprecipitation qPCR on AEP versus AT2 chromatin demonstrates Ctnnb1 antibody binding at the differentially accessible

genomic regions near Etv4, Sftpa, Lamp3 and Gpr116 in AEP cells, indicating that these genes are direct Wnt targets. Data are shown as mean with individual data points showing summary data from two independent chromatin immunoprecipitation experiments with multiple technical replicates. **f–j**, Fgfr2 activation in mouse AEPs drives increased proliferation and the formation of larger organoids; quantification shown in **j**. See Fig. 4 for additional data. **k**, RNAscope showing enriched expression of Fgfr2 (red) in lineage-labelled AEPs. **l–q**, Similar to treatment with Fgf7, Fgf10 treatment drives increased colony-forming efficiency in both mouse AEPs (**l–p**) and human AEPs (**q**). Data shown in **f–j**, **l–q** represent a minimum of  $n = 12$  wells across two individual experiments. Statistics are representative of all biological replicates. Data were analysed with ANOVA followed by preplanned pairwise comparisons and adjustment for multiple comparison testing, and are shown centred on mean with bars indicating standard deviation. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  and \*\*\*\* $P < 0.0001$ .



# The SMAD2/3 interactome reveals that TGF $\beta$ controls m<sup>6</sup>A mRNA methylation in pluripotency

Alessandro Bertero<sup>1†\*</sup>, Stephanie Brown<sup>1\*</sup>, Pedro Madrigal<sup>1,2</sup>, Anna Osnato<sup>1</sup>, Daniel Ortmann<sup>1</sup>, Loukia Yiangou<sup>1</sup>, Juned Kadiwala<sup>1</sup>, Nina C. Hubner<sup>3</sup>, Igor Ruiz de los Mozos<sup>4</sup>, Christoph Sadée<sup>4</sup>, An-Sofie Lenaerts<sup>1</sup>, Shota Nakanoh<sup>1</sup>, Rodrigo Grandy<sup>1</sup>, Edward Farnell<sup>5</sup>, Jernej Ule<sup>4</sup>, Hendrik G. Stunnenberg<sup>3</sup>, Sasha Mendjan<sup>1†</sup> & Ludovic Vallier<sup>1,2</sup>

**The TGF $\beta$  pathway has essential roles in embryonic development, organ homeostasis, tissue repair and disease<sup>1,2</sup>. These diverse effects are mediated through the intracellular effectors SMAD2 and SMAD3 (hereafter SMAD2/3), whose canonical function is to control the activity of target genes by interacting with transcriptional regulators<sup>3</sup>. Therefore, a complete description of the factors that interact with SMAD2/3 in a given cell type would have broad implications for many areas of cell biology. Here we describe the interactome of SMAD2/3 in human pluripotent stem cells. This analysis reveals that SMAD2/3 is involved in multiple molecular processes in addition to its role in transcription. In particular, we identify a functional interaction with the METTL3–METTL14–WTAP complex, which mediates the conversion of adenosine to N<sup>6</sup>-methyladenosine (m<sup>6</sup>A) on RNA<sup>4</sup>. We show that SMAD2/3 promotes binding of the m<sup>6</sup>A methyltransferase complex to a subset of transcripts involved in early cell fate decisions. This mechanism destabilizes specific SMAD2/3 transcriptional targets, including the pluripotency factor gene *NANOG*, priming them for rapid downregulation upon differentiation to enable timely exit from pluripotency. Collectively, these findings reveal the mechanism by which extracellular signalling can induce rapid cellular responses through regulation of the epitranscriptome. These aspects of TGF $\beta$  signalling could have far-reaching implications in many other cell types and in diseases such as cancer<sup>5</sup>.**

Activin and NODAL, two members of the TGF $\beta$  superfamily, play essential roles in cell fate decision in human pluripotent stem cells (hPSCs)<sup>6–8</sup>. Activin–NODAL signalling is necessary to maintain pluripotency, and inhibition of this pathway drives differentiation towards the neuroectoderm lineage<sup>6,9,10</sup>. Activin–NODAL signalling also cooperates with BMP and WNT pathways to drive mesendoderm specification<sup>11–14</sup>. We therefore used the differentiation of hPSCs into definitive endoderm as a model system to investigate the SMAD2/3 interactome during a dynamic cellular process. To allow a comprehensive and unbiased examination of the proteins that interact with SMAD2/3, we developed an optimized SMAD2/3 co-immunoprecipitation protocol that is compatible with mass-spectrometry analyses (Extended Data Fig. 1a, b and Supplementary Discussion). By examining human embryonic stem cells (hESCs) and hESCs that have been induced to differentiate to endoderm (Fig. 1a), we identified 89 interacting partners of SMAD2/3 (Fig. 1b, Extended Data Fig. 1c, d and Supplementary Table 1). Eleven of these proteins interacted with SMAD2/3 either in hESCs or the differentiating cells but not in both (Extended Data Fig. 1e), suggesting that the SMAD2/3 interactome is largely conserved across these two cell types (Supplementary Discussion). Notably, this list includes known SMAD2/3 transcriptional and epigenetic cofactors (including FOXH1, SMAD4, SNON, SKI, EP300, SETDB1 and CREBBP<sup>3</sup>). We also performed functional

experiments on FOXH1, EP300, CREBBP and SETDB1, which uncovered essential functions of these SMAD2/3 transcriptional and epigenetic cofactors in hPSC fate decisions (Extended Data Figs 2, 3 and Supplementary Discussion).

These proteomic experiments also show that SMAD2/3 interacts with complexes involved in functions that have, to our knowledge, not previously been associated with TGF $\beta$  signalling (Fig. 1b and Extended Data Fig. 1f), such as ERCC1–XPF and DAPK3–PAWR, which are involved in DNA repair and apoptosis, respectively. We also identified several factors involved in mRNA processing, modification and degradation (Fig. 1b), such as the METTL3–METTL14–WTAP complex (involved in deposition of N<sup>6</sup>-methyladenosine (m<sup>6</sup>A)), the PABP-dependent poly(A) nuclease complex (PAN, involved in mRNA decay), the cleavage factor complex CFIm (involved in pre-mRNA 3' end processing) and the NONO–SFPQ–PSPC1 factors (involved in RNA splicing and nuclear retention of defective RNAs). Overall, these results suggest that SMAD2/3 could be involved in a large number of biological processes in hPSCs, including not only transcriptional and epigenetic regulation, but also non-canonical functions of SMAD2/3.

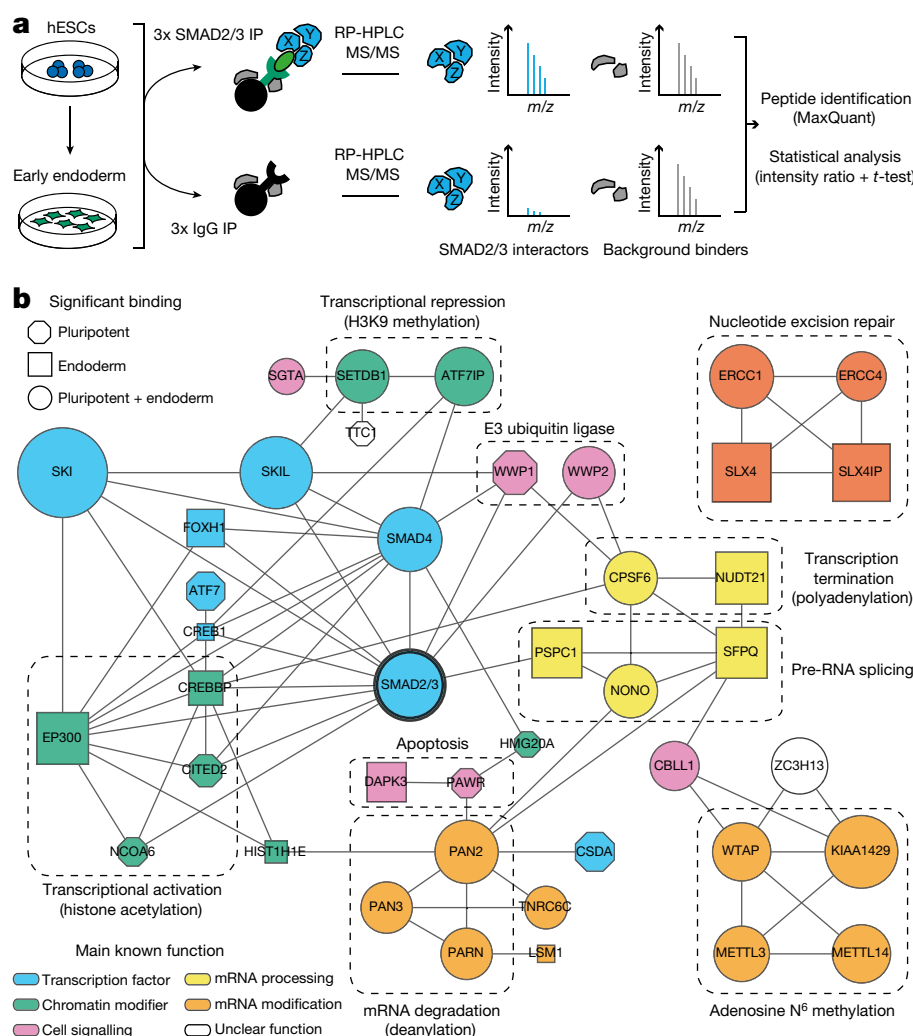
To further investigate these functions, we examined the function of activin–NODAL signalling in m<sup>6</sup>A deposition. m<sup>6</sup>A is the most common RNA modification, and it regulates multiple aspects of mRNA biology including decay and translation<sup>4,15–19</sup>. However, it is not known whether this is a dynamic event that can be modulated by extracellular cues. Furthermore, whereas m<sup>6</sup>A is known to regulate haematopoietic stem cells<sup>20,21</sup> and the transition between the naive and primed pluripotency states<sup>22,23</sup>, its function in hPSCs and during germ layer specification is unknown. We first validated the interaction of SMAD2/3 with METTL3–METTL14–WTAP by co-immunoprecipitation followed by western blotting with both hESCs and human induced pluripotent stem cells (hiPSCs; Fig. 2a and Extended Data Fig. 4a, b). Inhibition of SMAD2/3 phosphorylation blocked this interaction (Fig. 2b and Extended Data Fig. 4c). Proximity ligation assays (PLA) also demonstrated that the interaction occurs in the nucleus (Fig. 2c, d). These observations suggest that SMAD2/3 and the m<sup>6</sup>A methyltransferase complex interact and that this interaction depends on activin–NODAL signalling.

To investigate the functional relevance of this interaction, we assessed the transcriptome-wide effects of inhibition of activin–NODAL signalling on the deposition of m<sup>6</sup>A by performing nuclear-enriched m<sup>6</sup>A-methylated-RNA immunoprecipitation followed by deep sequencing (NeMeRIP-seq; Extended Data Fig. 5a–d, and Supplementary Discussion). Consistent with previous reports<sup>17,19,24</sup>, deposition of m<sup>6</sup>A onto exons was enriched around stop codons and transcription start sites, and occurred at a motif corresponding to the m<sup>6</sup>A-consensus sequence (Extended Data Fig. 5e–g). Assessment of differential m<sup>6</sup>A deposition revealed that inhibition of activin–NODAL

<sup>1</sup>Wellcome Trust–MRC Cambridge Stem Cell Institute, Anne McLaren Laboratory and Department of Surgery, University of Cambridge, Cambridge CB2 0SZ, UK. <sup>2</sup>Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. <sup>3</sup>Department of Molecular Biology, Radboud University, Nijmegen 6525GA, The Netherlands. <sup>4</sup>Francis Crick Institute and Department of Molecular Neuroscience, University College London, London NW1 1AT, UK. <sup>5</sup>Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK. <sup>†</sup>Present addresses: Department of Pathology, University of Washington, Seattle 98109, Washington, USA (A.B.); Institute of Molecular Biotechnology, Vienna 1030, Austria (S.M.)

\*These authors contributed equally to this work.





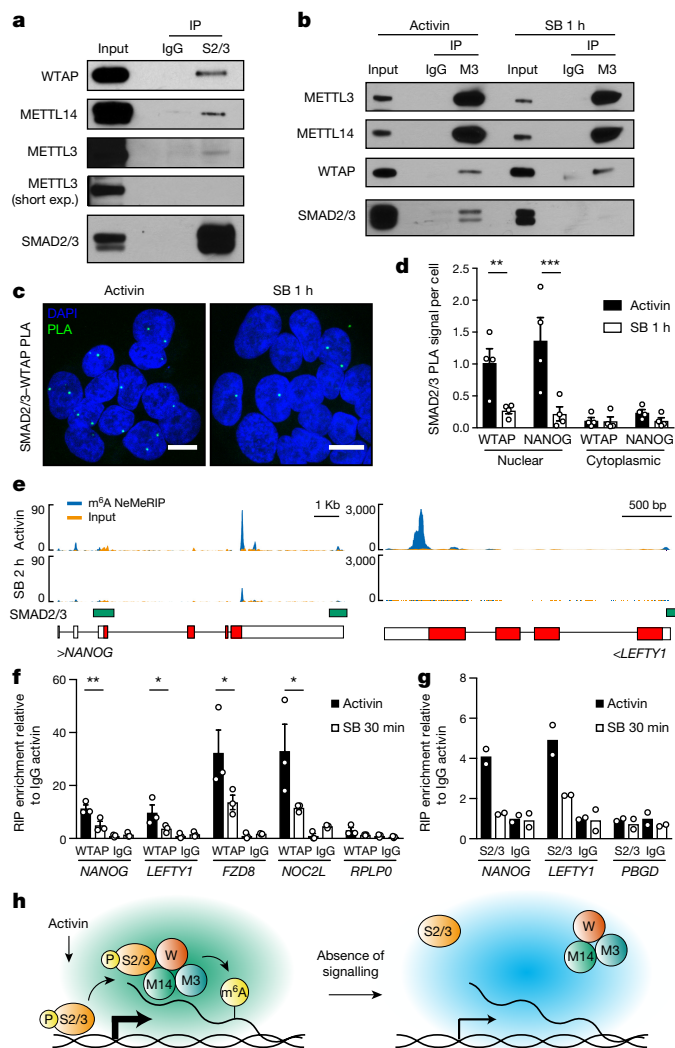
**Figure 1 | Identification of the SMAD2/3 interactome.** **a**, Experimental approach. IP, immunoprecipitation. RP-HPLC, reversed-phase high-performance liquid chromatography; MS/MS, tandem mass spectrometry. **b**, Interaction network of all known protein–protein interactions between selected SMAD2/3 partners identified in pluripotent and endoderm cells ( $n = 3$  co-immunoprecipitations; one-tailed  $t$ -test, permutation-based false discovery rate (FDR)  $< 0.05$ ). Nodes describe: (1) the lineage in which the proteins were significantly enriched (shape); (2) the significance of the enrichment (size is proportional to the maximum  $-\log P$  value); and (3) the function of the factors (colour). Complexes of interest are marked.

signalling predominantly resulted in decreased m<sup>6</sup>A levels in selected transcripts (Supplementary Table 2; mean absolute log<sub>2</sub> fold-change of 0.56 and 0.35 for m<sup>6</sup>A decrease and increase, respectively). Decreases in m<sup>6</sup>A deposition were mostly observed on peaks located near stop codons (Extended Data Fig. 5h), where m<sup>6</sup>A deposition has been reported to decrease the stability of mRNAs<sup>16,24,25</sup>. Transcripts with reduced m<sup>6</sup>A levels after inhibition of activin–NODAL signalling largely and significantly overlapped with genes bound by SMAD2/3 ( $P < 2.88 \times 10^{-18}$ ; Extended Data Fig. 5i), including well-known transcriptional targets such as *NANOG*, *NODAL*, *LEFTY1* and *SMAD7* (Fig. 2e and Extended Data Fig. 5j). Accordingly, activin–NODAL-sensitive m<sup>6</sup>A deposition was largely associated with transcripts that rapidly decreased in abundance during the exit from pluripotency triggered by inhibition of activin–NODAL signalling (Extended Data Fig. 6a). Transcripts that behaved in this fashion were enriched in pluripotency regulators and factors involved in the activin–NODAL signalling pathway (Supplementary Table 3). On the other hand, the expression of a large number of developmental regulators associated with activin–NODAL-sensitive m<sup>6</sup>A deposition remained unchanged following inhibition of activin–NODAL signalling (Extended Data Fig. 6a–c and Supplementary Table 3). Considered together, these findings show that activin–NODAL signalling can regulate m<sup>6</sup>A deposition on a number of specific transcripts.

We then examined the molecular mechanisms that underlie the regulation of m<sup>6</sup>A deposition by activin–NODAL signalling. RNA immunoprecipitation experiments on nuclear RNAs showed that inhibition of activin–NODAL signalling impaired binding of WTAP to several m<sup>6</sup>A-labelled transcripts including *NANOG* and *LEFTY1* (Fig. 2f and

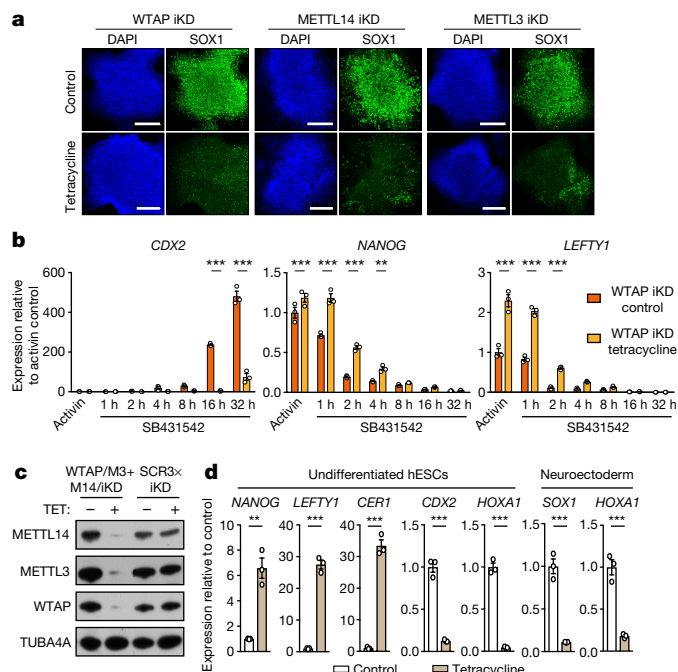
Extended Data Fig. 4d, e), whereas SMAD2/3 itself interacted with these transcripts in the presence of activin–NODAL signalling (Fig. 2g and Extended Data Fig. 4e). Thus, SMAD2/3 appears to promote the recruitment of the m<sup>6</sup>A methyltransferase complex to nuclear RNAs. Notably, recent reports have established that m<sup>6</sup>A deposition occurs co-transcriptionally and involves nascent pre-RNAs<sup>16,20,26</sup>. Considering the broad overlap between SMAD2/3 transcriptional targets and transcripts showing activin–NODAL-sensitive m<sup>6</sup>A deposition (Extended Data Fig. 5i), we hypothesized that SMAD2/3 could facilitate co-transcriptional recruitment of the m<sup>6</sup>A methyltransferase complex onto nascent transcripts. Consistent with this notion, inhibition of activin–NODAL signalling mainly resulted in downregulation of m<sup>6</sup>A, not only on exons, but also on pre-mRNA-specific features such as introns and exon–intron junctions (Extended Data Fig. 6d–i and Supplementary Table 2). Moreover, we observed a correlation in activin–NODAL sensitivity across m<sup>6</sup>A peaks within the same transcript (Extended Data Fig. 6j), suggesting that SMAD2/3 regulates m<sup>6</sup>A deposition at the level of the genomic locus rather than on a specific mRNA peak. Nevertheless, we did not detect stable and direct binding of the m<sup>6</sup>A methyltransferase complex to DNA (Extended Data Fig. 4f). Thus, co-transcriptional recruitment might rely on indirect and dynamic interactions with chromatin. Considering all these results, we propose a model in which activin–NODAL signalling promotes co-transcriptional m<sup>6</sup>A deposition by facilitating the recruitment of the m<sup>6</sup>A methyltransferase complex onto nascent mRNAs (Fig. 2h).

To understand the functional relevance of this regulation in the context of hPSC cell-fate decisions, we performed inducible knockdown of the subunits of the m<sup>6</sup>A methyltransferase complex<sup>27</sup> (Extended Data



**Figure 2 | Activin–NODAL signalling promotes m<sup>6</sup>A deposition on specific regulators of pluripotency and differentiation.** **a**, **b**, Western blots of SMAD2/3 (S2/3), METTL3 (M3) or control (IgG) immunoprecipitations from nuclear extracts of hESCs (representative of three experiments). Input is 5% of the material used for immunoprecipitation. In **b**, immunoprecipitations were performed on hESCs maintained in the presence of activin or treated for 1 h with the activin–NODAL inhibitor SB431542 (SB). For gel Source Data, see Supplementary Fig. 1. **c**, Proximity ligation assays (PLA) for SMAD2/3 and WTAP in hESCs maintained in the presence of activin or SB431542 (representative of two experiments). Scale bars, 10  $\mu$ m. DAPI, nuclei. **d**, PLA quantification; the known SMAD2/3 cofactor NANOG was used as positive control<sup>10</sup>. Mean  $\pm$  s.e.m.,  $n = 4$  PLAs; two-way analysis of variance (ANOVA) with post hoc Holm–Sidak comparisons. **e**, Representative results of nuclear-enriched m<sup>6</sup>A-methylated RNA immunoprecipitation followed by deep sequencing (m<sup>6</sup>A NeMeRIP-seq;  $n = 3$  cultures, replicates combined for visualization). Signal represents read enrichment normalized per million mapped reads and library size. GENCODE gene annotations (red, coding exons; white, untranslated exons; all potential exons are shown and overlaid) and SMAD2/3-binding sites from chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) data<sup>30</sup> are shown. **f**, **g**, RNA immunoprecipitation (RIP) experiments for WTAP, SMAD2/3 or IgG control in hESCs maintained in the presence of activin or treated with SB431542. *RPLP0* and *PBGD* were used as negative controls as they do not contain m<sup>6</sup>A. In **f**, mean  $\pm$  s.e.m.,  $n = 3$  cultures. Two-way ANOVA with post hoc Holm–Sidak comparisons. In **g**, bars show mean,  $n = 2$  cultures. **h**, Model of the mechanism by which SMAD2/3 promotes m<sup>6</sup>A deposition. P, phosphorylation; W, WTAP; M14, METTL14.

Fig. 7a, b). As expected, reducing expression of WTAP, METTL14 or METTL3 decreased m<sup>6</sup>A deposition (Extended Data Fig. 7c, d); however, prolonged knockdown did not affect pluripotency (Extended



**Figure 3 | The m<sup>6</sup>A methyltransferase complex antagonizes activin–NODAL signalling in hPSCs to promote timely exit from pluripotency.** **a**, Immunofluorescence for the neural marker SOX1 following neuroectoderm differentiation of tetracycline (TET)-inducible knockdown (iKD) hESCs (representative of two experiments). Control, no tetracycline; scale bars, 100  $\mu$ m. **b**, Quantitative PCR (qPCR) analyses in WTAP-iKD hESCs with inhibition of activin–NODAL signalling using SB431542 treatments with indicated duration. Mean  $\pm$  s.e.m.,  $n = 3$  cultures. Two-way ANOVA with post hoc Holm–Sidak comparisons. **c**, Western blot validation of multiple inducible knockdown (MiKD) hESCs (iKD of WTAP, METTL3 and METTL14). Cells expressing three copies of the scrambled shRNA (SCR3 $\times$ ) were used as negative control. **d**, qPCR analyses in undifferentiated MiKD hESCs, or following differentiation of MiKD hESCs to neuroectoderm. Mean  $\pm$  s.e.m.,  $n = 3$  cultures. Two-tailed *t*-test.

Data Fig. 7e, f). We also found that expression of m<sup>6</sup>A methyltransferase complex subunits was necessary for neuroectoderm differentiation induced by the inhibition of activin–NODAL signalling without being necessary for activin-driven endoderm specification (Fig. 3a and Extended Data Fig. 8a–c). Notably, Activin–NODAL is known to block neuroectoderm induction by promoting NANOG expression<sup>28</sup>, whereas NANOG is required for the early stages of endoderm specification<sup>13</sup>. Accordingly, we found that NANOG transcript and protein were upregulated, and the stability of NANOG mRNA increased when m<sup>6</sup>A methyltransferase activity was impaired (Fig. 3b and Extended Data Fig. 9a–c). These results show that m<sup>6</sup>A deposition decreases the stability of NANOG mRNA, facilitating its downregulation upon loss of activin–NODAL signalling, and thereby facilitating exit from pluripotency and neuroectoderm specification (Extended Data Fig. 9d). Additional transcriptomic analyses showed that WTAP knockdown resulted in global upregulation of genes that were transcriptionally activated by SMAD2/3 in hESCs and impaired the upregulation of genes induced by inhibition of activin–NODAL signalling during neuroectoderm differentiation (Fig. 3b, Extended Data Fig. 10a–e, Supplementary Table 4 and Supplementary Discussion). Notably, the decrease in WTAP expression also led to upregulation of m<sup>6</sup>A-marked mRNAs (Extended Data Fig. 10f), confirming that WTAP-dependent m<sup>6</sup>A deposition destabilizes mRNAs<sup>16,24,25</sup>. Moreover, transcripts that are rapidly downregulated after inhibition of activin–NODAL signalling were enriched in m<sup>6</sup>A-marked mRNAs (Extended Data Fig. 10f). Finally, simultaneous knockdown of METTL3, METTL14 and WTAP in hESCs resulted in an even stronger dysregulation of target transcripts of activin–NODAL signalling (Fig. 3c, d and Extended Data Fig. 8d) and

defective neuroectoderm differentiation (Fig. 3d and Extended Data Fig. 8e, f). Together, these results show that the interaction of SMAD2/3 with METTL3–METTL14–WTAP can promote m<sup>6</sup>A deposition on a subset of transcripts, including a number of pluripotency regulators that are also transcriptionally activated by activin–NODAL signalling. The resulting negative feedback destabilizes these mRNAs and causes their rapid degradation following inhibition of activin–NODAL signalling. This mechanism allows timely exit from pluripotency and induction of neuroectoderm differentiation (Extended Data Fig. 9d).

In conclusion, this analysis of the SMAD2/3 interactome reveals interactions between TGF $\beta$  signalling and a wide variety of cellular processes. Our results suggest that SMAD2/3 could act as a hub, coordinating several proteins known to have a role in mRNA processing and modification, apoptosis, DNA repair and transcriptional regulation. This function is illustrated by our results that show activin–NODAL-sensitive regulation of m<sup>6</sup>A. Activin–NODAL signalling connects transcriptional and epitranscriptional regulation through the interaction between SMAD2/3 and the METTL3–METTL14–WTAP complex, and primes its transcriptional targets for rapid degradation upon withdrawal of signalling (Extended Data Fig. 9d). This avoids overlaps between the pluripotency and neuroectoderm transcriptional programs, thereby facilitating changes in cell identity. We anticipate that further studies will clarify the other non-canonical functions of SMAD2/3, and will dissect how they are related to epigenetic, transcriptional and epitranscriptional regulation of gene expression.

Our findings also clarify and broaden our understanding of the function of m<sup>6</sup>A in cell-fate decisions. They establish that depletion of m<sup>6</sup>A in hPSCs does not lead to differentiation, contrary to predictions from studies in mouse-epiblast stem cells<sup>22</sup>. This could imply that there are important functional differences in epitranscriptional regulations between human and mouse pluripotent states. Moreover, widening the conclusions from previous reports<sup>23</sup>, we demonstrate that deposition of m<sup>6</sup>A is specifically necessary for neuroectoderm induction, but not for definitive endoderm differentiation. This can be explained by the fact that in contrast to its strong inhibitory effect on the neuroectoderm lineage<sup>28</sup>, expression of NANOG is necessary for the early stages of mesoderm specification<sup>13,29</sup>. Finally, our results establish that m<sup>6</sup>A modification of RNA is a dynamic event that is directly modulated by extracellular cues such as TGF $\beta$ . Considering the many functions of TGF $\beta$  signalling, the regulation we describe here may have an essential function in many cellular contexts that require a rapid response or change in cell state, such as the inflammatory response or cellular proliferation.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 29 November 2016; accepted 22 January 2018.**

**Published online 28 February 2018.**

- Wu, M. Y. & Hill, C. S. TGF $\beta$  superfamily signaling in embryonic development and homeostasis. *Dev. Cell* **16**, 329–343 (2009).
- Oshimori, N. & Fuchs, E. The harmonies played by TGF $\beta$  in stem cell biology. *Cell Stem Cell* **11**, 751–764 (2012).
- Gaarenstroom, T. & Hill, C. S. TGF $\beta$  signaling to chromatin: how SMADs regulate transcription during self-renewal and differentiation. *Semin. Cell Dev. Biol.* **32**, 107–118 (2014).
- Heyn, H. & Esteller, M. An adenine code for DNA: a second life for N<sup>6</sup>-methyladenine. *Cell* **161**, 710–713 (2015).
- Pickup, M., Novitskiy, S. & Moses, H. L. The roles of TGF $\beta$  in the tumour microenvironment. *Nat. Rev. Cancer* **13**, 788–799 (2013).
- Vallier, L., Reynolds, D. & Pedersen, R. A. NODAL inhibits differentiation of human embryonic stem cells along the neuroectodermal default pathway. *Dev. Biol.* **275**, 403–421 (2004).
- Vallier, L., Alexander, M. & Pedersen, R. A. Activin/NODAL and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *J. Cell Sci.* **118**, 4495–4509 (2005).
- James, D., Levine, A. J., Besser, D. & Hemmati-Brivanlou, A. TGF $\beta$ /activin/NODAL signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development* **132**, 1273–1282 (2005).
- Smith, J. R. et al. Inhibition of activin/NODAL signaling promotes specification of human embryonic stem cells into neuroectoderm. *Dev. Biol.* **313**, 107–117 (2008).

- Bertero, A. et al. Activin/NODAL signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. *Genes Dev.* **29**, 702–717 (2015).
- D'Amour, K. A. et al. Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat. Biotechnol.* **23**, 1534–1541 (2005).
- Vallier, L. et al. Signaling pathways controlling pluripotency and early cell fate decisions of human induced pluripotent stem cells. *Stem Cells* **27**, 2655–2666 (2009).
- Teo, A. K. et al. Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev.* **25**, 238–250 (2011).
- Kubo, A. et al. Development of definitive endoderm from embryonic stem cells in culture. *Development* **131**, 1651–1662 (2004).
- Ke, S. et al. A majority of m<sup>6</sup>A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev.* **29**, 2037–2053 (2015).
- Ke, S. et al. m<sup>6</sup>A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes Dev.* **31**, 990–1006 (2017).
- Dominissini, D. et al. Topology of the human and mouse m<sup>6</sup>A RNA methylomes revealed by m<sup>6</sup>A-seq. *Nature* **485**, 201–206 (2012).
- Meyer, K. D. et al. 5' UTR m<sup>6</sup>A promotes cap-independent translation. *Cell* **163**, 999–1010 (2015).
- Meyer, K. D. et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**, 1635–1646 (2012).
- Barbieri, I. et al. Promoter-bound METTL3 maintains myeloid leukaemia by m<sup>6</sup>A-dependent translation control. *Nature* **552**, 126–131 (2017).
- Vu, L. P. et al. The N<sup>6</sup>-methyladenosine (m<sup>6</sup>A)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nat. Med.* **23**, 1369–1376 (2017).
- Geula, S. et al. Stem cells. m<sup>6</sup>A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. *Science* **347**, 1002–1006 (2015).
- Batista, P. J. et al. m<sup>6</sup>A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* **15**, 707–719 (2014).
- Schwartz, S. et al. Perturbation of m<sup>6</sup>A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. *Cell Reports* **8**, 284–296 (2014).
- Wang, X. et al. N<sup>6</sup>-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117–120 (2014).
- Bartosovic, M. et al. N<sup>6</sup>-methyladenosine demethylase FTO targets pre-mRNAs and regulates alternative splicing and 3'-end processing. *Nucleic Acids Res.* **45**, 11356–11370 (2017).
- Bertero, A. et al. Optimized inducible shRNA and CRISPR/Cas9 platforms for *in vitro* studies of human development using hPSCs. *Development* **143**, 4405–4418 (2016).
- Vallier, L. et al. Activin/NODAL signalling maintains pluripotency by controlling NANOG expression. *Development* **136**, 1339–1349 (2009).
- Mendjan, S. et al. NANOG and CDX2 pattern distinct subtypes of human mesoderm during exit from pluripotency. *Cell Stem Cell* **15**, 310–325 (2014).
- Brown, S. et al. Activin/NODAL signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors. *Stem Cells* **29**, 1176–1185 (2011).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank Cambridge Genomic Services for help with next-generation sequencing. This work was supported by the European Research Council starting grant 'Relieve IMDs' (L.V., S.B., A.B., P.M.); the Cambridge University Hospitals National Institute for Health Research Biomedical Research Center (L.V., J.K., A.-S.L.); the Wellcome Trust PhD program (A.O., L.Y.); a British Heart Foundation PhD studentship (FS/11/77/39327 to A.B.); a Grant-in-Aid for JSPS Fellows (16J08005 to S.N.); and a core support grant from the Wellcome Trust and Medical Research Council to the Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute.

**Author Contributions** A.B. conceived the study, performed or contributed to most of the experiments, analysed data and wrote the manuscript with input from the other authors. S.B. contributed to study conception, performed co-immunoprecipitation, NeMeRIP and RNA-IP experiments, and analysed data. P.M., I.R.d.I.M. and C.S. analysed NeMeRIP-seq. A.O. performed PLA and co-immunoprecipitations and analysed RNA-seq data. D.O., L.Y. and J.K. assisted with hPSC gene editing and differentiation; N.C.H. performed quantitative proteomics and data analysis. A.-S.L., S.N. and R.G. assisted with hPSC culture. E.F. optimized NeMeRIP-seq libraries. J.U. contributed to study conception and supervision. H.G.S. supervised quantitative proteomics. S.M. contributed to study conception and supervision, and assisted with SMAD2/3 co-immunoprecipitation. L.V. conceived, supervised and supported the study, and wrote and provided final approval of the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to L.V. (lv225@cam.ac.uk).

**Reviewer Information** Nature thanks C. Mason and the other anonymous reviewer(s) for their contribution to the peer review of this work.



## METHODS

**hPSC culture and differentiation.** Feeder- and serum-free culture of hESCs (H9/WA09 line; WiCell) and hiPSCs (A1AT<sup>R/R</sup>; ref. 31) have been previously described<sup>32</sup>. In brief, cells were plated on gelatin and MEF medium-coated plates, and cultured in chemically defined medium (CDM) containing bovine serum albumin (BSA). CDM was supplemented with 10 ng/ml activin-A and 12 ng/ml FGF2 (both from M. Hyvonen, Dept. of Biochemistry, University of Cambridge). Cells were passaged every 5–6 days with collagenase IV and plated as clumps of 50–100 cells dispensed at a density of 100–150 clumps per cm<sup>2</sup>. Differentiation was initiated in adherent hESC cultures 48 h after passaging. Definitive endoderm specification was induced for three days (unless stated otherwise) by culturing cells in CDM (without insulin) with 20 ng/ml FGF2, 10  $\mu$ M LY294002 (PI3K inhibitor; Promega), 100 ng/ml activin-A and 10 ng/ml BMP4 (R&D), as previously described<sup>33</sup>. Neuroectoderm was induced for three days (unless stated otherwise) in CDM-BSA with 12 ng/ml FGF2 and 10  $\mu$ M SB431542 (activin-NODAL-TGF $\beta$  signalling inhibitor; Tocris), as previously described<sup>34</sup>. These same culture conditions were used for activin-NODAL signalling inhibition experiments. hPSCs were routinely monitored for absence of karyotypic abnormalities and mycoplasma infection. As hESCs were obtained from a commercial supplier, cell line identification was not performed. hiPSCs were previously generated in house and genotyped by Sanger sequencing<sup>31</sup>.

**Molecular cloning.** Plasmids carrying inducible shRNAs were generated by cloning annealed oligonucleotides into the pAAV-Puro\_iKD or pAAV-Puro\_siKD vectors as previously described<sup>27</sup>. All shRNA sequences were obtained from the RNAi Consortium TRC library<sup>35</sup> (<https://www.broadinstitute.org/rnai/public/>). Of the shRNAs that had been validated, the most powerful ones were chosen (the sequences are reported in Supplementary Table 5). Generation of a vector containing shRNAs against METTL3, METTL14 and WTAP (cloned in this order) was performed by Gibson assembly of PCR products containing individual shRNA cassettes, as previously described<sup>27</sup>. The resulting vector was named pAAV-Puro\_MsiKD-M3M14W. Generation of the matched control vector containing three copies of the scrambled shRNA sequence (pAAV-Puro\_MsiKD-SCR3 $\times$ ) has been described previously<sup>27</sup>.

A targeting vector for the *AAVS1* locus carrying constitutively-expressed *NANOG* was generated starting from pAAV\_TRE-eGFP<sup>36</sup>. First, the TRE-eGFP cassette was removed using PspXI and EcoRI, and substituted with the CAG promoter (cut from pR26-CAG-eGFP<sup>27</sup> using SpeI and BamHI) by ligating blunt-ended fragments. The resulting vector (pAAV-Puro\_CAG) was then used to clone the full-length *NANOG* transcript, which includes its full 5' and 3' UTRs. The full-length *NANOG* transcript was constructed from three DNA fragments. The 5' (bases 1–301) and 3' (bases 1878–2105) ends were synthesized (IDT) with 40 bp overlaps corresponding to pGem3Z vector linearized with SmaI. The middle fragment was amplified from cDNA of H9 hESCs obtained by retrotranscription with poly-T primer using primers 5'-TTGTCCCCAAAGCTTGCTT-3' and 5'-CAAAAACGGTAAGAAATCAATTAA-3'. The three fragments and the linearized vector were assembled using a Gibson reaction (NEB) and the sequence of the construct was confirmed by Sanger sequencing. The full length *NANOG* transcript was then subcloned into KpnI- and EcoRV-digested pAAV-Puro\_CAG following KpnI and HincII digestion. The resulting vector was named pAAV-Puro\_CAG-NANOG.

**Inducible gene knockdown.** Clonal inducible knockdown hESCs for METTL3, METTL14, WTAP or matched controls expressing a scrambled (SCR) shRNA were generated by gene targeting of the *AAVS1* locus with pAAV-Puro\_siKD plasmids, which was verified by genomic PCR, all as previously described<sup>27,36</sup>. This same approach was followed to generate multiple inducible knockdown hESCs for METTL3, METTL14 and WTAP (plasmid pAAV-Puro\_MsiKD-M3M14W) or matched controls expressing three copies of the SCR shRNA (plasmid pAAV-Puro\_MsiKD-SCR3 $\times$ ). Inducible knockdown hESCs for SMAD2, FOXH1, SETDB1, EP300, CREBBP, B2M and matched controls expressing a scrambled shRNA were generated using pAAV-Puro\_iKD vectors<sup>27</sup> in hESCs expressing a randomly integrated wild-type tetracycline resistance gene. Two wells were transfected for each shRNA in order to generate independent biological replicates. Following selection with puromycin, the resulting targeted cells in each well were pooled and expanded for further analysis. Given that 20 to 50 clones were obtained for each well, we refer to these lines as 'clonal pools'. Gene knockdown was induced by adding 1  $\mu$ g/ml tetracycline hydrochloride (Sigma-Aldrich) to the culture medium. Unless indicated otherwise in the text or figure legends, inducible knockdown in undifferentiated hESCs was induced for five days, while differentiation assays were performed in hESCs in which knockdown had been induced for ten days.

**Generation of NANOG-overexpressing hESCs.** NANOG-overexpressing H9 hESCs were obtained by zinc-finger nuclease (ZFN)-facilitated gene targeting of the *AAVS1* locus with pAAV-Puro\_CAG-NANOG. This was performed by lipofection of the targeting vector and zinc-finger plasmids followed by puromycin

selection, clonal isolation and genotyping screening of targeted cells, all as previously described<sup>27</sup>.

**SMAD2/3 co-immunoprecipitation.** Approximately  $2 \times 10^7$  cells were used for each immunoprecipitation. Unless stated otherwise, all biochemical steps were performed on ice or at 4°C, and ice-cold buffers were supplemented with complete protease inhibitors (Roche), PhosSTOP Phosphatase Inhibitor Cocktail (Roche), 1 mg/ml leupeptin, 0.2 mM DTT, 0.2 mM PMSF and 10 mM sodium butyrate (all from Sigma-Aldrich). Cells were fed with fresh medium for 2 h before being washed with PBS, scraped in cell dissociation buffer (CDB, Gibco) and pelleted at 250 g for 10 min. The cell pellet was then washed once with 10 volumes of PBS, and once with 10 volumes of hypotonic lysis buffer (HLB; 10 mM HEPES pH 7.6, 10 mM KCl, 2 mM MgCl<sub>2</sub>, 0.2 mM EDTA, 0.2 mM EGTA). The pellet was resuspended in 5 volumes of HLB and incubated for 5 min to induce cell swelling. The resulting cell suspension was homogenized using the 'loose' pestle of a Dounce homogenizer (Jencons Scientific) for 35–50 strokes until plasma membrane lysis was complete (as judged by microscopic inspection). The nuclei were pelleted at 800 g for 5 min, washed once with ten volumes of HLB, and resuspended in 1.5 volumes of high-salt nuclear lysis buffer (HSNLB; 20 mM HEPES pH 7.6, 420 mM NaCl, 2 mM MgCl<sub>2</sub>, 25% glycerol, 0.2 mM EDTA, 0.2 mM EGTA). High-salt nuclear extraction was performed by homogenizing the nuclei using the 'tight' pestle of a Dounce homogenizer for 70 strokes, followed by 45 min of incubation with rotation. The resulting lysate was clarified for 30 min at 16,000g and transferred to a dialysis cassette using a 19-gauge syringe. Dialysis was performed for 4 h in 1 l of dialysis buffer (20 mM HEPES pH 7.6, 50 mM KCl, 100 mM NaCl, 2 mM MgCl<sub>2</sub>, 10% glycerol, 0.2 mM EDTA, 0.2 mM EGTA) with gentle stirring, and the buffer was changed once after 2 h. After dialysis, the sample was clarified from minor protein precipitates for 10 min at 17,000g, and the protein concentration was assessed. Immunoprecipitations were performed by incubating 0.5 mg of protein with 5  $\mu$ g of goat polyclonal SMAD2/3 antibody (R&D systems, AF3797) or goat IgG negative control antibody (R&D systems, AB-108-C) for 3 h at 4°C with rotation. This was followed by incubation with 10  $\mu$ l of protein-G agarose for 1 h. Beads were washed three times with dialysis buffer and processed for western blot or mass spectrometry. This co-immunoprecipitation protocol is referred to as 'co-IP2' in the Supplementary Discussion and in Extended Data Fig. 1. The alternative SMAD2/3 co-immunoprecipitation protocol (co-IP1) has been previously described<sup>10</sup>.

**Mass spectrometry.** Label-free quantitative mass-spectrometric analysis of proteins co-immunoprecipitated with SMAD2/3 or from control IgG co-immunoprecipitations was performed on three replicates for each condition. After immunoprecipitation, samples were prepared as previously described<sup>37</sup> with minor modifications. Proteins were eluted by incubation with 50  $\mu$ l of 2 M urea and 10 mM DTT for 30 min at room temperature with agitation. Then, 55 mM chloroacetamide was added for 20 min to alkylate reduced disulfide bonds. Proteins were pre-digested on the beads with 0.4  $\mu$ g of mass-spectrometry-quality trypsin (Promega) for 1 h at room temperature with agitation. The suspension was cleared from the beads by centrifugation. The beads were then washed with 50  $\mu$ l of 2 M urea, and the combined supernatants were incubated overnight at room temperature with agitation to complete digestion. 0.1% trifluoroacetic acid was then added to inactivate trypsin, and peptides were loaded on C<sub>18</sub> StageTips<sup>38</sup>. Tips were prepared for binding by sequential equilibration for 2 min at 800g with 50  $\mu$ l methanol, 50  $\mu$ l Solvent B (0.5% acetic acid; 80% acetonitrile) and 50  $\mu$ l Solvent A (0.5% acetic acid). Subsequently, peptides were loaded and washed twice with Solvent A. Tips were stored in dry conditions until analysis. Peptides were eluted from the StageTips and separated by reversed-phase liquid chromatography on a 2.5-h-long segmented gradient using EASY-nLC 1000 (ThermoFisher Scientific). Eluting peptides were ionized and injected directly into a Q Exactive mass spectrometer (ThermoFisher Scientific). The mass spectrometer was operated in TOP10 sequencing mode, meaning that one full mass-spectrometry scan was followed by higher energy collision induced dissociation (HCD) and subsequent detection of the fragmentation spectra of the 10 most abundant peptide ions (MS/MS). Collectively, ~160,000 isotype patterns were generated resulting from ~6,000 mass-spectrometry runs. Consequently, ~33,000 MS/MS spectra were measured.

Quantitative mass spectrometry based on dimethyl labelling of samples was performed as described for label-free quantitative mass spectrometry but with the following differences. Dimethyl labelling was performed as previously reported<sup>39,40</sup>. In brief, trypsin-digested protein samples were incubated with dimethyl labelling reagents (4  $\mu$ l of 0.6 M NaBH<sub>3</sub>CN together with 4  $\mu$ l of 4% CH<sub>2</sub>O or CD<sub>2</sub>O for light or heavy labelling, respectively) for 1 h at room temperature with agitation. The reaction was stopped by adding 16  $\mu$ l of 1% NH<sub>3</sub>. Samples were acidified with 0.1% trifluoroacetic acid, and finally loaded on StageTips. Each immunoprecipitation was performed twice, switching the labels.

**Analysis of mass-spectrometry data.** The raw label-free quantitative mass-spectrometry data were analysed using the MaxQuant software suite<sup>41</sup>.



Peptide spectra were compared against the human database (Uniprot) using the integrated Andromeda search engine, and peptides were identified with  $FDR < 0.01$ , determined by false matches against a reverse decoy database. Peptides were assembled into protein groups with an  $FDR < 0.01$ . Protein quantification was performed using the MaxQuant label-free quantification algorithm requiring at least two ratio counts, in order to obtain label-free quantification (LFQ) intensities. Collectively, the MS/MS spectra were matched to ~20,000 known peptides, leading to the identification of 3,635 proteins in at least one of the conditions analysed. Statistical analysis of the data was performed using the Perseus software package (MaxQuant). First, common contaminants and reverse hits were removed, and only proteins identified by at least two peptides (one of those being unique to the respective protein group) were considered as high-confidence identifications. Proteins were then filtered for those identified in all replicates of at least one condition. LFQ intensities were converted to their log values, and missing intensity values were imputed by representative noise values<sup>42</sup>. One-tailed *t*-tests were then performed to determine the specific interactors in each condition by comparing the immunoprecipitations with the SMAD2/3 antibody to those with the IgG negative controls. Statistical significance was set with a permutation-based  $FDR < 0.05$  (250 permutations). Fold-enrichment over IgG controls was calculated from LFQ intensities.

This same pipeline was used to analyse mass-spectrometry data based on dimethyl labelling, with the following two exceptions. First, an additional mass of 28.03 Da (light) or 32.06 Da (heavy) was specified as 'labels' at the N terminus and at lysines. Second, during statistical analysis of mass-spectrometry data, the outlier significance was calculated based on protein intensity (significance  $B^{41}$ ), and was required to be below 0.05 for both the forward and the reverse experiments.

**Biological interpretation of mass-spectrometry data.** The SMAD2/3 protein–protein interaction network was generated using Cytoscape v.2.8.3<sup>43</sup>. First, all the annotated interactions involving the SMAD2/3-binding proteins were inferred by interrogating protein–protein interaction databases through the PSIQUIC Universal Web Service Client. IMEX-complying interactions were retained and merged by union. Then, a subnetwork involving only the SMAD2/3 interactors was isolated. Finally, duplicate nodes and self-loops were removed to simplify visualization. Note that based on our results all the proteins shown would be connected to SMAD2/3, but such links were omitted to simplify visualization and highlight those interactions with SMAD2/3 that were already known. Proteins lacking any link and small complexes of less than three factors were not shown, in order to improve presentation clarity. Note that since the nodes representing SMAD2 and SMAD3 shared the same links, they were fused into a single node (SMAD2/3). Functional enrichment analysis was performed using Fisher's exact test implemented in Enrichr<sup>44</sup>, and only enriched terms with a Benjamini–Hochberg adjusted *P* value  $< 0.05$  were considered. For Gene Ontology (GO) enrichment analysis, the 2015 GO annotation was used. For mouse phenotype enrichment analysis, level 3 of the Mouse Genomic Informatics (MGI) annotation was used. To compare protein abundance in different conditions, a cut-off of absolute LFQ intensity  $\log_2$  fold-change larger than 2 was chosen, as label-free mass spectrometry is currently not sensitive enough to detect smaller changes with confidence<sup>37</sup>.

**Proximity ligation assay.** Proximity ligation assay (PLA) was performed using the Duolink *In situ* Red Starter Kit Goat/Rabbit (Sigma-Aldrich). Cells were cultured on glass coverslips and prepared by fixation in 4% paraformaldehyde (PFA) in PBS for 10 min at room temperature, followed by two gentle washes in PBS. All subsequent incubations were performed at room temperature unless otherwise stated. Samples were permeabilized in PBS containing 0.25% Triton X-100 for 20 min, blocked in PBS with 0.5% BSA for 30 min, and incubated with the two primary antibodies of interest (diluted in PBS with 0.5% BSA; see Supplementary Table 6) for 1 h at 37 °C in a humid chamber. The Duolink *In situ* PLA probes (anti-rabbit minus and anti-goat plus) were mixed and diluted 1:5 in PBS with 0.5% BSA, and pre-incubated for 20 min. Following two washes with PBS containing 0.5% BSA, the coverslips were incubated with the PLA probe solution for 1 h at 37 °C in a humidified chamber. Single-antibody and probes-only negative controls were performed for each antibody tested to confirm assay specificity. Coverslips were washed twice in wash buffer A for 5 min under gentle agitation, and incubated with 1× ligation solution supplemented with DNA ligase (1:40 dilution) for 30 min at 37 °C in a humidified chamber. After two more washes in wash buffer A for 2 min with gentle agitation, coverslips were incubated with 1× amplification solution supplemented with DNA polymerase (1:80 dilution) for 1 h 40 min at 37 °C in a humid chamber. Samples were protected from light from this step onwards. Following two washes in wash buffer B for 10 min, the coverslips were dried overnight, and finally mounted on a microscope slide using Duolink *In situ* Mounting Medium with DAPI. Images of random fields of view were acquired using a LSM 700 confocal microscope (Leica) using a Plan-Apochromat 40×/1.3 Oil DIC M27 objective, performing z-stack with optimal spacing (~0.36 μm). Images were analysed automatically using ImageJ. For this, nuclear (DAPI) and

PLA z-stacks were first individually flattened (max intensity projection) and thresholded to remove background noise. Nuclear images were further segmented using the watershed function. Total nuclei and PLA spots were quantified using the 'analyse particle' function of ImageJ, and nuclear PLA spots were quantified using the 'speckle inspector' function of the ImageJ plugin BioVoxsel.

**RNA immunoprecipitation.** Approximately  $2 \times 10^7$  cells were used for each RNA immunoprecipitation (RIP). Unless stated otherwise, all biochemical steps were performed on ice or at 4 °C, and ice-cold buffers were supplemented with cComplete Protease Inhibitors (Roche) and PhosSTOP Phosphatase Inhibitor Cocktail (Roche). Cells were fed with fresh culture medium 2 h before being washed once with room-temperature PBS and UV crosslinked in PBS at room temperature using a Stratilinker 1800 at 254 nm (400 mJ/cm<sup>2</sup>). Crosslinked cells were scraped off in cell-dissociation buffer (CDB, Gibco) and pelleted at 250g for 5 min. The cell pellet was incubated in five volumes of isotonic lysis buffer (ILB; 10 mM Tris-HCl pH 7.5, 3 mM CaCl<sub>2</sub>, 2 mM MgCl<sub>2</sub>, 0.32 M sucrose) for 12 min to induce cell swelling. Then, Triton X-100 was added to a final concentration of 0.3%, and cells were incubated for 6 min to lyse the plasma membranes. Nuclei were pelleted at 600g for 5 min, washed once with ten volumes of ILB, and finally resuspended in two volumes of nuclear lysis buffer (NLB; 50 mM Tris-HCl pH 7.5, 100 mM NaCl, 50 mM KCl, 3 mM MgCl<sub>2</sub>, 1 mM EDTA, 10% glycerol, 0.1% Tween) supplemented with 800 U/ml RNasin Ribonuclease Plus Inhibitor (Promega) and 1 μM DTT. The nuclear suspension was transferred to a Dounce homogenizer (Jencons Scientific) and homogenized by performing 70 strokes with a tight pestle. The nuclear lysate was incubated with rotation for 30 min, homogenized again by performing 30 additional strokes with the tight pestle, and incubated in rotation for 15 min at room temperature after addition of 12.5 μg/ml of DNase I (Sigma). The protein concentration was assessed, and approximately 1 mg of protein was used for overnight immunoprecipitation with rotation with the primary antibody of interest (Supplementary Table 6), or with equal amounts of non-immune species-matched IgG. Ten per cent of the protein lysate used for immunoprecipitation was saved as pre-immunoprecipitation input and stored at –80 °C for subsequent RNA extraction. Immunoprecipitation reactions were then incubated for 1 h with 30 μl of protein-G agarose, then washed twice with 1 ml of LiCl wash buffer (50 mM Tris-HCl pH 7.5, 250 mM LiCl, 0.1% Triton X-100, 1 mM DTT) and twice with 1 ml of NLB. Beads were resuspended in 90 μl of 30 mM Tris-HCl pH 9.0, and DNase-digested using the RNase-free DNase kit (QIAGEN) by adding 10 μl of RDD buffer and 2.5 μl of DNase. The pre-immunoprecipitation input samples were similarly treated in parallel, and samples were incubated for 10 min at room temperature. The reaction was stopped by adding 2 mM EDTA and by heating at 70 °C for 5 min. Proteins were digested by adding 2 μl of proteinase K (20 mg/ml; Sigma-Aldrich) and by incubating at 37 °C for 30 min. Finally, RNA was extracted with 1 ml of TriReagent (Sigma-Aldrich) according to the supplier's instructions. The RNA was resuspended in nuclease-free water, and half of the sample was used in a reverse-transcription reaction using SuperScript II (ThermoFisher) using the manufacturer's protocol. The other half was used in a control reaction with no reverse transcriptase to confirm successful removal of DNA contaminants. Samples were quantified by quantitative real-time PCR (qPCR), and normalized first to the pre-immunoprecipitation input and then to the IgG control using the  $\Delta\Delta C_t$  approach (see below). Supplementary Table 5 shows the primers used.

**Chromatin immunoprecipitation.** Approximately  $2 \times 10^7$  cells were used for each chromatin immunoprecipitation (ChIP), and cells were fed with fresh medium 2 h before collection. ChIP was performed using a previously described protocol<sup>10,30</sup>. Briefly, cells were crosslinked on plates, first with protein–protein crosslinkers (10 mM dimethyl 3,3'-dithiopropionimide dihydrochloride and 2.5 mM 3,3'-dithiodipropionic acid di-*N*-hydroxysuccinimide ester; Sigma-Aldrich) for 15 min at room temperature, then with 1% formaldehyde for 15 min. Crosslinking was quenched with glycine, after which cells were collected, subjected to nuclear extraction, and sonicated to fragment the DNA. Following pre-clearing, the lysate was incubated overnight with the antibodies of interest (Supplementary Table 6) or non-immune IgG. ChIP was completed by incubation with protein-G agarose beads followed by subsequent washes with high salt and LiCl-containing buffers (all exactly as previously described<sup>10,30</sup>). Crosslinking was reverted, first by adding DTT (for disulfide bridge-containing protein–protein crosslinkers), then by incubating in high salt at high temperature. DNA was finally purified by sequential phenol–chloroform and chloroform extractions. Samples were analysed by qPCR using the  $\Delta\Delta C_t$  approach (see Supplementary Table 5 for primer sequences). First, a region in the last exon of *SMAD7* was used as internal control to normalize for background binding. Second, the enrichment was normalized to the enrichment observed in non-immune IgG ChIP controls.

**m<sup>6</sup>A dot blot.** m<sup>6</sup>A dot blots were performed as described with minor modifications<sup>23</sup>. Poly-A RNA was purified from total cellular RNA using the Dynabeads mRNA Purification Kit (ThermoFisher), diluted in 50 μl of RNA loading buffer (RLB; 2.2 M formaldehyde, 50% formamide, 0.5× MOPS buffer

(20 mM MOPS, 12.5 mM CH<sub>3</sub>COONa, 1.25 mM EDTA, pH 7.0)), incubated at 55°C for 15 min, and snap-cooled on ice. An Amersham Hybond-XL membrane was rehydrated in water for 3 min, then in 10× saline-sodium citrate buffer (SSC; 1.5 M NaCl, 150 mM Na<sub>2</sub>C<sub>6</sub>H<sub>5</sub>O<sub>7</sub>, pH 7.0) for 10 min, and finally ‘sandwiched’ in a 96-well dot blot hybridization manifold (ThermoFisher Scientific). Following two washes of the wells with 150 µl of 10× SSC, the RNA was spotted onto the membrane. After UV crosslinking for 2 min at 254 nm using a Stratilinker 1800 (Stratagene), the membrane was washed once with Tris-buffered saline Tween buffer (TBST; 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1% Tween-20), and blocked for 1 h at room temperature with TBST supplemented with 4% non-fat dry milk. Incubations with the anti-m<sup>6</sup>A primary antibody (Synaptic Systems, 202-111; used at 1 µg/ml) and the mouse-HRP secondary antibody (Supplementary Table 6) were each performed in TBST with 4% milk for 1 h at room temperature, and were followed by three 10-min washes at room temperature in TBST. Finally, the membrane was incubated with ECL2 Western Blotting Substrate (Pierce), and exposed to X-Ray Super RX film.

**m<sup>6</sup>A nuclear-enriched methylated RNA immunoprecipitation.** m<sup>6</sup>A MeRIP of nuclear-enriched RNA for analysis by deep sequencing (NeMeRIP-seq) was performed using modifications of previously described methods<sup>23,45</sup>. For each sample,  $7.5 \times 10^7$  hESCs were used, and three biological replicates were performed per condition. Cells were fed with fresh medium for 2 h before washing with PBS, scraping in cell dissociation buffer (CDB, Gibco), and pelleting at 250 g for 5 min. The cell pellet was incubated in five volumes of isotonic lysis buffer (ILB; 10 mM Tris-HCl pH 7.5, 3 mM CaCl<sub>2</sub>, 2 mM MgCl<sub>2</sub>, 0.32 M sucrose, 1,000 U/ml RNasin RNase inhibitor (Promega), 1 mM DTT) for 10 min to induce cell swelling. Then, Triton X-100 was added to a final concentration of 0.3% and cells were incubated for 6 min to lyse the plasma membranes. Nuclei were pelleted at 600 g for 5 min and washed once with ten volumes of ILB. RNA was extracted from the nuclear pellet using the RNeasy midi kit (QIAGEN) according to the manufacturer’s instructions. Residual contaminating DNA was digested in solution using the RNase-free DNase set from QIAGEN, and RNA was re-purified by sequential acid phenol–chloroform and chloroform extractions followed by ethanol precipitation. At this stage, complete removal of DNA contamination was confirmed by qPCR of the resulting RNA without a reverse-transcription step. RNA was then chemically fragmented in 20 µl reactions each containing 20 µg of RNA in fragmentation buffer (10 mM ZnCl<sub>2</sub>, 10 mM Tris-HCl pH 7.0). Such reactions were incubated at 95°C for 5 min, followed by inactivation with 50 mM EDTA and storage on ice. The fragmented RNA was then cleaned up by ethanol precipitation. In preparation for MeRIP, 15 µg of anti-m<sup>6</sup>A antibody (Synaptic Systems, 202-003) or equivalent amounts of rabbit non-immune IgG were crosslinked to 0.5 mg of magnetic beads using the Dynabeads Antibody Coupling Kit (ThermoFisher Scientific) according to the manufacturer’s instructions. Following equilibration of the magnetic beads by washing with 500 µl of binding buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1% NP-40, 1 mM EDTA), MeRIP reactions were assembled with 300 µg of the fragmented RNA in 3 ml of binding buffer supplemented with 3,000 U of RNasin RNase inhibitor. Samples were incubated with rotation at 7 r.p.m. for 1 h at room temperature. Fragmented RNA (5 µg, 10% of the amount used for MeRIP) was set aside as pre-MeRIP input control. MeRIP reactions were washed twice with binding buffer, once with low-salt buffer (LSB; 0.25× SSPE (saline-sodium phosphate-EDTA buffer; 150 mM NaCl, 10 mM NaHPO<sub>4</sub>, 10 mM Na<sub>2</sub>EDTA, pH 7.4), 37.5 mM NaCl, 1 mM EDTA, 0.05% Tween-20), once with high-salt buffer (HSB; 0.25× SSPE, 137.5 mM NaCl, 1 mM EDTA, 0.05% Tween-20), and twice with TE–Tween buffer (TTB; 10 mM Tris-HCl pH 7.4, 1 mM EDTA, 0.05% Tween-20). Each wash was performed by incubating the beads with 500 µl of buffer at 7 r.p.m. for 3 min at room temperature. Finally, RNA was eluted from the beads by four successive incubations with 75 µl of elution buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 20 mM DTT, 0.1% SDS, 1 mM EDTA) at 42°C. Both the RNA from pooled MeRIP eluates and the pre-MeRIP input were purified and concentrated by sequential acid phenol–chloroform and chloroform extractions followed by ethanol precipitation. Glycogen (30 µg) was added as carrier during ethanol precipitation. RNA was resuspended in 15 µl of ultrapure RNase-free water. DNA libraries were prepared for deep sequencing using the TruSeq Stranded total RNA kit (Illumina) according to the manufacturer’s instructions with the following exceptions: (1) Ribo-Zero treatment was performed only for pre-NeMeRIP samples, as there was minimal ribosomal RNA contamination in m<sup>6</sup>A NeMeRIP samples; (2) since samples were pre-fragmented, the fragmentation step was bypassed and 30 ng of RNA from each sample was used directly for library prep; (3) owing to the small size of the library, a twofold excess of Ampure XP beads was used during all purification steps in order to retain small fragments; (4) owing to the presence of contaminating adapter dimers, the library was gel-extracted using gel-safe stain and a dark reader in order to remove fragments smaller than ~120 bp. Pooled libraries were diluted and denatured for sequencing on the NextSeq 500 (Illumina) according to the manufacturer’s instructions. Samples were pooled to obtain >30 million

unique clusters per sample. The PhiX control library (Illumina) was spiked into the main library pool at 1% vol/vol for quality control purposes. Sequencing was performed using a high output flow cell with  $2 \times 75$  cycles of sequencing, which provided ~800 million paired-end reads from ~400 million unique clusters from each lane. Overall, an average of ~33 million and ~54 million paired-end reads were generated for m<sup>6</sup>A MeRIP and pre-MeRIP samples, respectively.

m<sup>6</sup>A MeRIP samples to be analysed by qPCR (NeMeRIP-qPCR) were processed as described for NeMeRIP-seq, but starting from  $2.5 \times 10^7$  cells. MeRIP of cytoplasmic RNA was performed using RNA extracted from the cytoplasmic fraction of cells that were being processed for NeMeRIP. In both cases, MeRIP was performed as for NeMeRIP-seq, but using 2.5 µg of anti-m<sup>6</sup>A antibody (or equivalent amounts of rabbit non-immune IgG) and 50 µg RNA in 500 µl binding buffer supplemented with 500 U RNasin RNase inhibitor. At the end of the protocol, the RNA was resuspended in 15 µl ultrapure RNase-free water. For m<sup>6</sup>A MeRIP on total RNA, the protocol just described was followed exactly, with the exception that the subcellular fractionation step was bypassed, and that total RNA was extracted from  $5 \times 10^6$  cells. For m<sup>6</sup>A MeRIP on mRNA, poly-A RNA was purified from 75 µg total RNA using the Dynabeads mRNA Purification Kit, and 2.5 µg of the resulting mRNA was used for chemical fragmentation and subsequent MeRIP with 1 µg anti-m<sup>6</sup>A antibody. At the end of all these protocols, cDNA synthesis was performed using all of the MeRIP material in a 30 µl reaction containing 500 ng random primers, 0.5 mM dNTPs, 20 U RNaseOUT, and 200 U SuperScript II (all from Invitrogen) according to the manufacturer’s instructions. cDNA was diluted tenfold, and 5 µl of this dilution was used for qPCR using KAPA Sybr Fast Low Rox (KAPA Biosystems). For each gene of interest, two primer pairs were designed against either the region containing the m<sup>6</sup>A peak<sup>23</sup> or against a negative region (a portion of the same transcript lacking the m<sup>6</sup>A peak; Supplementary Table 5). Results of MeRIP-qPCR for each gene were then calculated using the  $\Delta\Delta C_t$  approach by using the negative region to normalize both for the expression level of the transcript of interest and for background binding.

**Analysis of NeMeRIP-seq data.** Quality of raw sequencing data was assessed using Trimmomatic v.0.35<sup>46</sup>, with parameters ‘LEADING:3 TRAILING:3 SLIDINGWINDOW:5:10 MINLEN:40’. Reads were aligned to the GRCh38 human genome assembly using TopHat 2.0.13<sup>47</sup> with parameters ‘-library-type fr-firststrand -transcriptome-index’ and the Ensembl GRCh38.83 annotation. Identification of novel splice junctions was allowed. Paired-end and unpaired reads passing quality control were concatenated and mapped in ‘single-end’ mode in order to be used with MetDiff<sup>48</sup>, which supports only single-end reads. Reads with MAPQ < 20 were filtered out. m<sup>6</sup>A peak calling and differential RNA methylation in the exome were assessed using MetDiff<sup>48</sup> with pooled inputs for each condition, ‘GENE\_ANNO\_GTF = GRCh38.83, MINIMAL\_MAPQ = 20’, and of the remaining parameters as default (PEAK\_CUTOFF\_FDR = 0.05; DIFF\_PEAK\_CUTOFF\_FDR = 0.05). MetDiff calculates *P* values using a likelihood ratio test, then adjusts them to FDR by Benjamini–Hochberg correction. An additional cut-off of absolute fold-change > 1.5 (meaning an absolute log<sub>2</sub> fold-change > 0.585) was applied for certain analyses as specified in the figure legends or tables. Given known differences between epitranscriptome maps as a function of pipeline<sup>49,50</sup>, we confirmed the site-specific and general trends in our data by using an additional pipeline<sup>45</sup>. For this, MACS2<sup>51</sup> was used with parameters ‘-q 0.05–nomodel–keep-dup all’ in m<sup>6</sup>A NeMeRIP-seq and paired inputs after read alignment with Bowtie 2.2.2.0 (reads with MAPQ < 20 were filtered out). Peaks found in at least two samples were kept for further processing, and a consensus MACS2 peak list was obtained, merging those located within a distance less than 100 bp. The MetDiff and MACS2 peak lists largely overlapped (Extended Data Fig. 5d), and differed primarily because MACS2 identifies peaks throughout the genome while MetDiff identifies only peaks found on the exome (Extended Data Fig. 5c). For the following analyses focused on exonic m<sup>6</sup>A peaks, we considered a stringent consensus list of only those MetDiff peaks that overlapped with MACS2 peaks (Supplementary Table 2, ‘exon m6a’). We assessed the reproducibility of m<sup>6</sup>A NeMeRIP-seq triplicates in peak regions using the Bioconductor package fCCAC v1.0.0<sup>52</sup>. Hierarchical clustering (Euclidean distance, complete method) of *F* values corresponding to the first two canonical correlations divided the samples into activin and SB431542 clusters. Normalized read-coverage files were generated using the function ‘normalize\_bigwig’ in RSeQC-2.6<sup>53</sup> with default parameters. The distribution of m<sup>6</sup>A coverage across genomic features was plotted using the Bioconductor package RCAS<sup>54</sup> with sampleN = 0 (no downsampling) and flankSize = 2500. Motif finding on m<sup>6</sup>A peaks was performed using DREME with default parameters<sup>55</sup>. For visualization purposes, the three biological replicates were combined. The Biodalliance genome viewer<sup>56</sup> was used to generate figures. Gene expression in this experiment was estimated from the pre-MeRIP input samples (which represent an RNA-seq sample on nuclear-enriched RNA species). Quantification, normalization of read counts and estimation of differential gene expression in pre-MeRIP input samples were performed using featureCounts<sup>57</sup>



and DESeq2<sup>58</sup>. For assessment of reproducibility, regularized log transformation of count data was computed, and biological replicates of input samples of the same condition were clustered together in the PC space<sup>59</sup>. Estimation of differential m<sup>6</sup>A deposition onto each peak in NeMeRIP samples versus input controls was performed using an analogous approach. Functional-enrichment analysis of m<sup>6</sup>A-marked transcripts was performed using Enrichr<sup>44</sup>, as described above for mass-spectrometry data. The coordinates of SMAD2/3 ChIP-seq peaks in hESCs<sup>30</sup> were transferred from their original mappings on hg18 to hg38 using liftOver. Overlap of the resulting intervals with m<sup>6</sup>A peaks significantly downregulated after 2 h of SB431542 treatment was determined using GAT<sup>60</sup> with default parameters. SMAD2/3-binding sites were assigned to the nearest gene using the annotatePeaks.pl function from the HOMER suite<sup>61</sup> with standard parameters. The significance in the overlap between the resulting gene list and that of genes encoding for transcripts with m<sup>6</sup>A peaks that are significantly downregulated after 2 h of SB431542 treatment was calculated by a hypergeometric test where the population size corresponded to the number of genes in the standard Ensembl annotation (GRCh38.83).

m<sup>6</sup>A peaks on introns were identified in three steps (Extended Data Fig. 6d). First, MetDiff was used to simultaneously perform peak calling and differential methylation analysis. Since MetDiff only accepts a transcriptome GTF annotation as an input to determine the genomic space onto which it identifies m<sup>6</sup>A peaks, in order to determine peaks on introns, we followed the strategy recommended by the package developers of running the software using a custom transcriptome annotation that includes introns<sup>48,62</sup>. This 'extended' transcriptome annotation was built using Cufflinks 2.2.1<sup>63</sup> with parameters '-library-type=fr-firststrand -m 100 -s 50' and guided by the Ensembl annotation (GRCh38.83). This was assembled using all available pre-NeMeRIP input reads. The result was an extended transcriptome annotation including all of the transcribed genome that could be detected and reconstructed from our nuclear-enriched input RNA samples, thus including most expressed introns. Then, MetDiff was run using this extended annotation as input for GENE\_ANNO\_GTF, pooled inputs for each condition, WINDOW\_WIDTH = 40, SLIDING\_STEP = 20, FRAGMENT\_LENGTH = 250, PEAK\_CUTOFF\_PVALUE = 1E-03, FOLD\_ENRICHMENT = 2, MINIMAL\_MAPQ = 20, and all other parameters as default. In a second step, the peaks identified by MetDiff were filtered for robustness by requiring that they overlapped with MACS2 peak calls, exactly as for exome-focused MetDiff peak calls (Extended Data Fig. 5d). Finally, only peaks that strictly did not overlap with any exon based on the Human Gencode annotation V.27 were retained to ensure specificity of mapping to introns (Supplementary Table 2; 'intron m<sup>6</sup>A'). MetDiff scores for the resulting peak list were used to assess differential m<sup>6</sup>A deposition based on the cutoff of FDR < 0.05.

m<sup>6</sup>A exon peaks spanning splice sites were selected from those identified by both the MetDiff analysis on the transcribed genome that was just described and by MACS2. Among these peaks, those presenting sequencing reads that overlap both an exon and an upstream or downstream intron were further selected (Supplementary Table 2; 'splice-site spanning m<sup>6</sup>A'). Peaks accomplishing MetDiff-calculated FDR < 0.05 and absolute fold-change > 1.5 (log<sub>2</sub> fold-change < -0.585) were used to create densities of RPKM-normalized reads inside exons and in the ±500 bp surrounding introns. Biological replicates were merged and depicted on 10 bp-binned heatmaps for visualization purposes. To study the covariation of m<sup>6</sup>A peaks inside each transcriptional unit, the exonic peak with the greatest downregulated MetDiff fold-change was compared to the mean fold-change of the rest of the m<sup>6</sup>A peaks found within the gene (both on exons and on introns). The resulting correlation was significant ( $P < 2 \times 10^{-16}$ ; adjusted  $R^2 = 0.2221$ ).

**RNA sequencing.** Poly-A purified opposing-strand-specific mRNA libraries were prepared from 200 ng of total RNA using the TruSeq Stranded mRNA HT sample preparation kit (Illumina). Samples were individually indexed for pooling using a dual-index strategy. Libraries were quantified both with a Qubit (ThermoFisher Scientific) and by qPCR using the NGS Library Quantification Kit (KAPA Biosystems). Libraries were then normalized and pooled. Pooled libraries were diluted and denatured for sequencing on the NextSeq 500 (Illumina) according to the manufacturer's instructions. Samples were pooled so as to obtain >30 million unique clusters per sample (18 samples were split in two runs and multiplexed across four lanes per run). The PhiX control library (Illumina) was spiked into the main library pool at 1% vol/vol for quality control purposes. Sequencing was performed using a high-output flow cell with  $2 \times 75$  cycles of sequencing, which provided ~800 million paired-end reads from ~400 million unique clusters from each run. Overall, a total of ~80 million paired-end reads per sample were obtained.

**Analysis of RNA-seq data.** Reads were trimmed using Sickle<sup>64</sup> with ' $q = 20$  and ' $l = 30$ '. To prepare for reads alignment, the human transcriptome was built with TopHat2 v.2.1.0<sup>44</sup> based on Bowtie v.2.2.6<sup>65</sup> by using the human GRCh38.p6 as reference genome, and the Ensembl gene transfer format (GTF) as annotation ([http://ftp.ensembl.org/pub/release-83/gtf/homo\\_sapiens/](http://ftp.ensembl.org/pub/release-83/gtf/homo_sapiens/)). All analyses were performed using this transcriptome assembly. Alignment was performed using

TopHat2 with standard parameters. Using Samtools view<sup>66</sup>, reads with MAPQ > 10 were kept for further analyses. Subsequent quantitative data analysis was performed using SeqMonk<sup>67</sup>. The RNA-seq pipeline was used to quantify gene expression as reads per million mapped reads (RPM), and differential expression analysis for binary comparisons was performed using the R package DESeq2<sup>58</sup>. A combined cut-off of negative binomial test  $P < 0.05$  and abs.FC > 2 was chosen. Analysis of differentially expressed transcripts across all samples was done using the Bioconductor timecourse package<sup>68</sup> in R. The Hotelling  $T^2$  score for each transcript was calculated using the MB.2D function with all parameters set to their default value. Hotelling  $T^2$  scores were used to rank probes according to differential expression across the time course, and the top 5% differentially expressed transcripts were selected for complete Euclidean hierarchical clustering ( $k$ -means preprocessing; max of 300 clusters) using Perseus software. Z-scores of log<sub>2</sub> normalized expression values across the timecourse were calculated and used for this analysis. Eight gene clusters were defined, and gene-enrichment analysis for selected clusters was performed using the Fisher's exact test implemented in Enrichr<sup>44</sup>. Only enriched terms with a Benjamini-Hochberg adjusted  $P$  value < 0.05 were considered. Principal component analysis was performed on the same list of top 5% differentially expressed transcripts using Perseus.

**Quantitative real-time PCR.** Cellular RNA was extracted using the GenElute Mammalian Total RNA Miniprep Kit and the On-Column DNase I Digestion Set (both from Sigma-Aldrich) following the manufacturer's instructions. 500 ng of RNA was used for complementary DNA (cDNA) synthesis using SuperScript II (Invitrogen) according to the manufacturer's instructions. cDNA was diluted 30-fold, and 5 µl was used for qPCR with SensiMix SYBR low-ROX (Bioline) and 150 nM forward and reverse primers (Sigma-Aldrich; see Supplementary Table 5 for primer sequences). Samples were run as technical duplicates in 96-well plates on a Stratagene Mx-3005P (Agilent), and results were analysed using the delta-delta cycle threshold ( $\Delta\Delta C_t$ ) approach<sup>69</sup> using RPLP0 as housekeeping gene. The reference sample used as control to calculate the relative gene expression is indicated in each figure or figure legend. In cases where multiple control samples were used as reference, the average  $\Delta C_t$  from all controls was used when calculating the  $\Delta\Delta C_t$ . All primers were designed using PrimerBlast (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>), and were validated to have a qPCR efficiency >98% and to produce a single PCR product.

**mRNA-stability measurements.** RNA stability was measured by collecting RNA samples at different time points following transcriptional inhibition with 1 µg/ml actinomycin D (Sigma-Aldrich). Following qPCR analyses using equal amounts of mRNA, gene expression was expressed as relative to the beginning of the experiment (no actinomycin D treatment). The data were then fitted to a one-phase decay-regression model<sup>70</sup>, and statistical differences in mRNA half-life were evaluated by comparing the model fits by extra sum-of-squares  $F$  test.

**Western blots.** Samples were prepared by adding Laemmli buffer (final concentrations: 30 mM Tris-HCl pH 6.8, 6% glycerol, 2% SDS, 0.02% bromophenol blue and 0.25% β-mercaptoethanol), and were denatured at 95 °C for 5 min. Proteins were loaded and run on 4–12% NuPAGE Bis-Tris Precast Gels (Invitrogen), then transferred to polyvinylidene fluoride (PVDF) membranes by liquid transfer using NuPAGE Transfer buffer (Invitrogen). Membranes were blocked for 1 h at room temperature in PBS, 0.05% Tween-20 (PBST) supplemented with 4% non-fat dried milk, and incubated overnight at 4 °C with primary antibody diluted in the same blocking buffer (Supplementary Table 6). After three washes in PBST, membranes were incubated for 1 h at room temperature with horseradish peroxidase (HRP)-conjugated secondary antibodies diluted in blocking buffer (Supplementary Table 6), then washed a further three times with PBST before incubation with Pierce ECL2 Western Blotting Substrate (Thermo) and exposure with X-Ray Super RX Films (Fujifilm).

**Immunofluorescence.** Cells were fixed for 20 min at 4 °C in PBS with 4% PFA, rinsed three times with PBS, and blocked and permeabilized for 30 min at room temperature using PBS with 10% donkey serum (Biorad) and 0.1% Triton X-100 (Sigma-Aldrich). Primary antibodies (Supplementary Table 6) were diluted in PBS with 1% donkey serum and 0.1% Triton X-100 and incubated overnight at 4 °C. This was followed by three washes with PBS and further incubation with AlexaFluor secondary antibodies (Supplementary Table 6) for 1 h at room temperature away from light. Cells were finally washed three times with PBS, and DAPI (Sigma-Aldrich) was added to the first wash to stain nuclei. Images were acquired using a LSM 700 confocal microscope (Leica).

**Flow cytometry.** Single-cell suspensions were prepared by incubation in cell-cell dissociation buffer (CDB; Gibco) for 10 min at 37 °C followed by extensive pipetting. Cells were washed twice with PBS and fixed for 20 min at 4 °C with PBS, 4% PFA. After three washes with PBS, cells were first permeabilized for 20 min at room temperature with PBS, 0.1% Triton X-100, then blocked for 30 min at room temperature with PBS containing 10% donkey serum. Primary and secondary antibody incubations (Supplementary Table 6) were performed for 1 h each at

room temperature in PBS, 1% donkey serum, 0.1% Triton X-100, and cells were washed three times with this same buffer after each incubation. Flow cytometry was performed using a Cyan ADP flow cytometer, and at least 10,000 events were recorded. Data analysis was performed using FlowJo X.

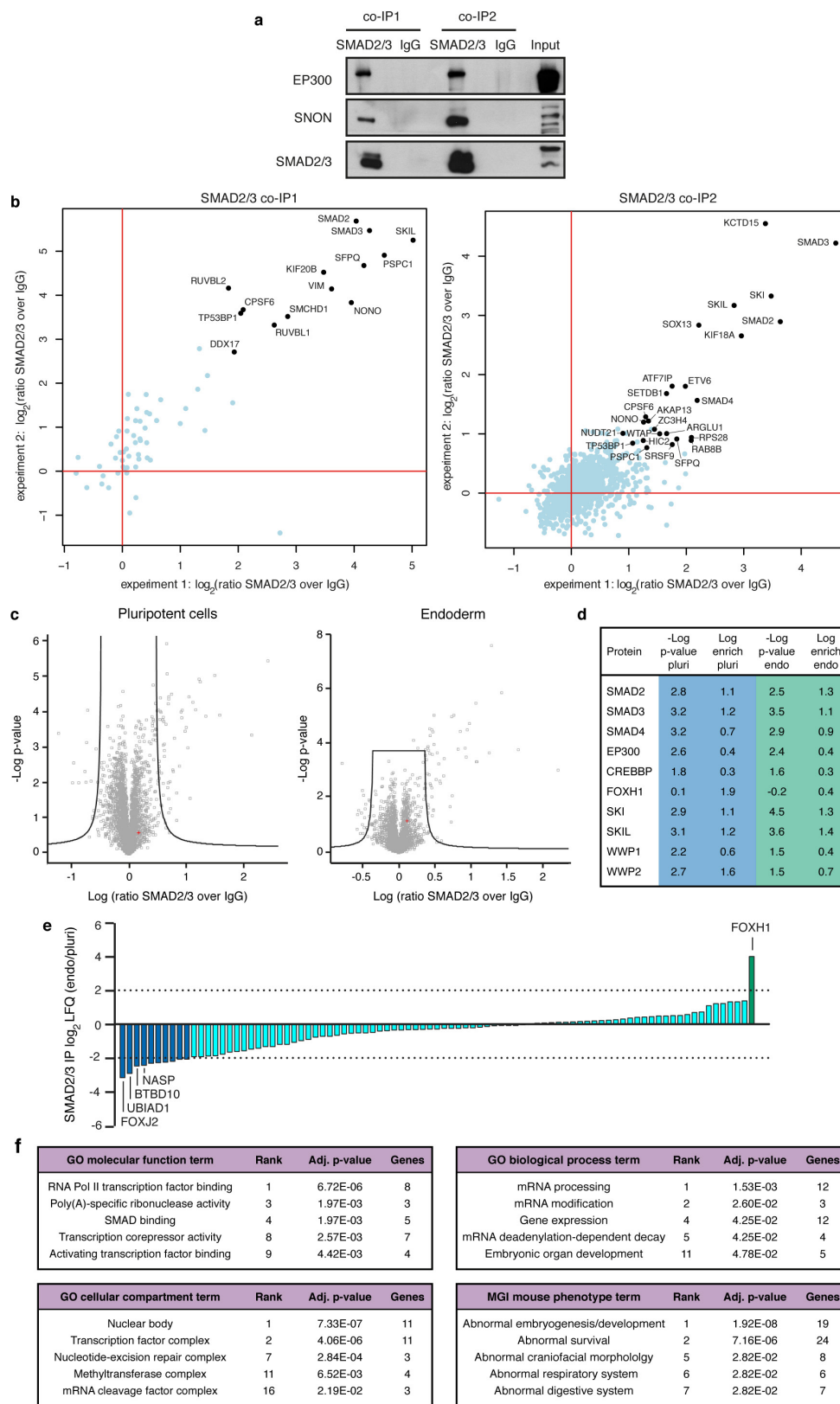
**Statistics and reproducibility.** Unless described otherwise in a specific section of the Methods, standard statistical analyses were performed using GraphPad Prism 7 using default parameters. The type and number of replicates, the statistical test used, and the test results are described in the figure legends. The level of significance in all graphs is represented as follows: \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$ . Test assumptions (for example, normal distribution) were confirmed where appropriate. For analyses with  $n < 10$ , individual data points are shown, and the mean  $\pm$  s.e.m. is reported for all analyses with  $n > 2$ . The mean is reported when  $n = 2$ , and no other statistics were calculated for these experiments owing to the small sample size. No experimental samples were excluded from the statistical analyses. Sample size was not pre-determined through power calculations, and no randomization or investigator blinding approaches were implemented during the experiments and data analyses. When representative results are presented, the experiments were reproduced in at least two independent cultures, and the exact number of such replications is detailed in the figure legend.

**Code availability.** Custom bioinformatics scripts used to analyse the data presented in the study have been deposited in GitHub (<http://github.com/pmb59/neMeRIP-seq>).

**Data availability.** The mass-spectrometry proteomics data that support the findings of this study have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the identifier PXD005285. Nucleotide sequencing data that support the findings of this study have been deposited to Array Express with identifiers E-MTAB-5229 and E-MTAB-5230. Source Data for the graphical representations found in all figures and Extended Data figures are provided in the Supplementary Information of this manuscript. Electrophoretic gel Source Data (uncropped scans with size marker indications) are presented in Supplementary Fig. 1. Supplementary Tables 1 to 4 provide the results of bioinformatics analyses described in the text and figure legends. All other data that support the findings of this study are available from the corresponding author upon reasonable request.

31. Yusa, K. *et al.* Targeted gene correction of  $\alpha 1$ -antitrypsin deficiency in induced pluripotent stem cells. *Nature* **478**, 391–394 (2011).
32. Vallier, L. Serum-free and feeder-free culture conditions for human embryonic stem cells. *Methods Mol. Biol.* **690**, 57–66 (2011).
33. Touboul, T. *et al.* Generation of functional hepatocytes from human embryonic stem cells under chemically defined conditions that recapitulate liver development. *Hepatology* **51**, 1754–1765 (2010).
34. Vallier, L. *et al.* Early cell fate decisions of human embryonic stem cells and mouse epiblast stem cells are controlled by the same signalling pathways. *PLoS One* **4**, e6082 (2009).
35. Moffat, J. *et al.* A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283–1298 (2006).
36. Pawlowski, M. *et al.* Inducible and deterministic forward programming of human pluripotent stem cells into neurons, skeletal myocytes, and oligodendrocytes. *Stem Cell Reports* **8**, 803–812 (2017).
37. Hubner, N. C. & Mann, M. Extracting gene function from protein-protein interactions using quantitative BAC interactomics (QUBIC). *Methods* **53**, 453–459 (2011).
38. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
39. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).
40. Hubner, N. C., Nguyen, L. N., Hornig, N. C. & Stunnenberg, H. G. A quantitative proteomics tool to identify DNA-protein interactions in primary cells or blood. *J. Proteome Res.* **14**, 1315–1329 (2015).
41. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
42. Hubner, N. C. *et al.* Quantitative proteomics combined with BAC transgenomics reveals in vivo protein interactions. *J. Cell Biol.* **189**, 739–754 (2010).
43. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
44. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
45. Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N. & Rechavi, G. Transcriptome-wide mapping of N<sup>6</sup>-methyladenosine by m<sup>6</sup>A-seq based on immunocapturing and massively parallel sequencing. *Nat. Protoc.* **8**, 176–189 (2013).
46. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
47. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
48. Cui, X. *et al.* MeTDiff: a novel differential RNA methylation analysis for MeRIP-seq data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **PP**, 1 (2015).
49. Saleatore, Y. *et al.* The birth of the epitranscriptome: deciphering the function of RNA modifications. *Genome Biol.* **13**, 175 (2012).
50. Li, X., Xiong, X. & Yi, C. Epitranscriptome sequencing technologies: decoding RNA modifications. *Nat. Methods* **14**, 23–31 (2017).
51. Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
52. Madrigal, P. fCCAC: functional canonical correlation analysis to evaluate covariance between nucleic acid sequencing datasets. *Bioinformatics* **33**, 746–748 (2017).
53. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
54. Uyar, B. *et al.* RCAS: an RNA centric annotation system for transcriptome-wide regions of interest. *Nucleic Acids Res.* **45**, e91 (2017).
55. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).
56. Down, T. A., Piipari, M. & Hubbard, T. J. P. Dalliace: interactive genome viewing on the web. *Bioinformatics* **27**, 889–890 (2011).
57. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
58. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
59. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
60. Heger, A., Webber, C., Goodson, M., Ponting, C. P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046–2048 (2013).
61. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
62. Meng, J., Cui, X., Rao, M. K., Chen, Y. & Huang, Y. Exome-based analysis for RNA epigenome sequencing data. *Bioinformatics* **29**, 1565–1567 (2013).
63. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
64. Joshi, N. & Fass, J. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files v.1.33 <https://github.com/najoshi/sickle> (2011).
65. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
66. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
67. Andrews, S. SeqMonk: A tool to visualise and analyse high throughput mapped sequence data <https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/> (2014).
68. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, e3 (2004).
69. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2<sup>- $\Delta\Delta C_t$</sup>  method. *Methods* **25**, 402–408 (2001).
70. Harrold, S., Genovese, C., Kobrin, B., Morrison, S. L. & Milcarek, C. A comparison of apparent mRNA half-life using kinetic labeling techniques vs decay following administration of transcriptional inhibitors. *Anal. Biochem.* **198**, 19–29 (1991).



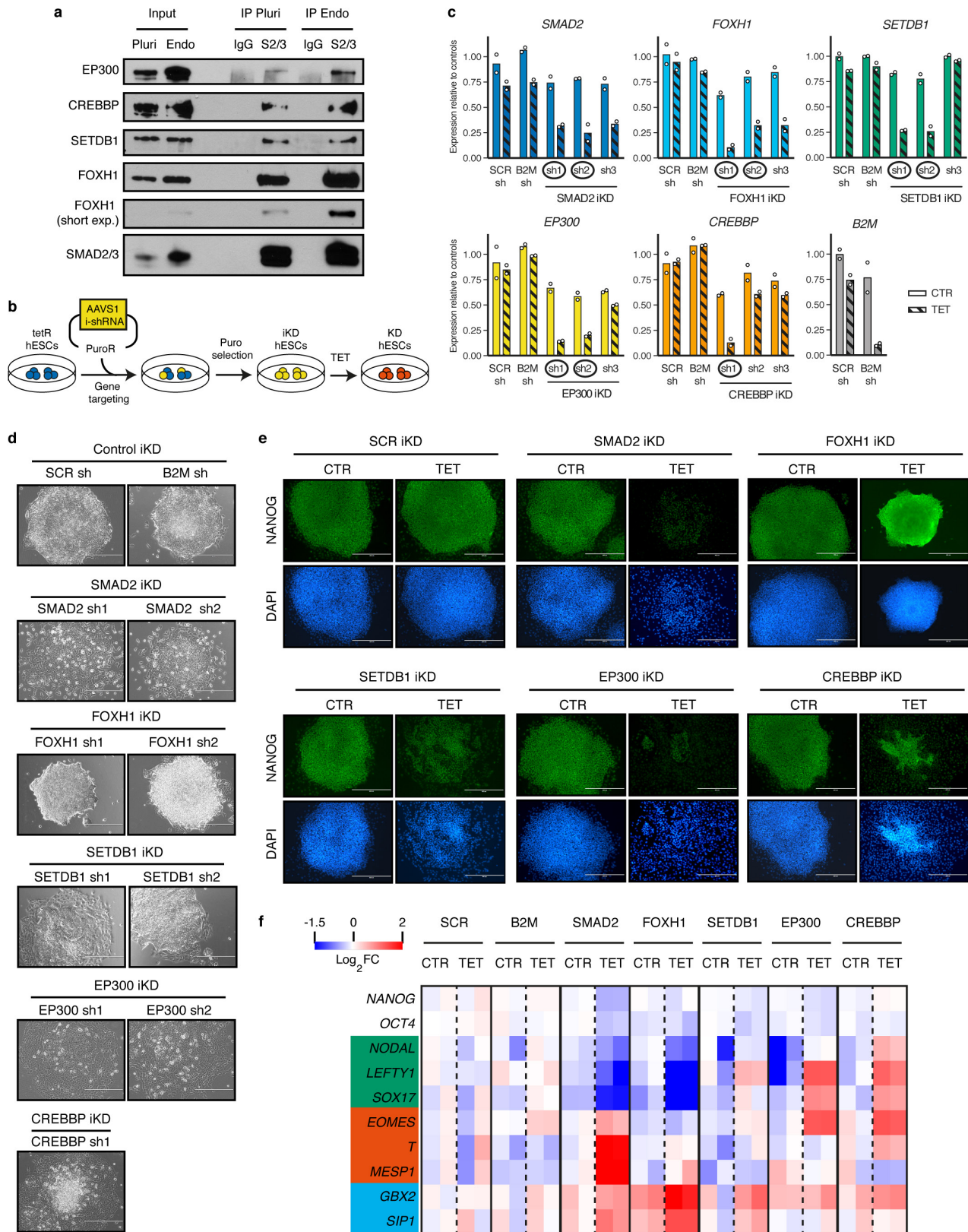


Extended Data Figure 1 | See next page for caption.

### Extended Data Figure 1 | Optimized co-immunoprecipitation protocol to define the SMAD2/3 interactome in hPSCs and early endoderm cells.

**a**, Western blots of SMAD2/3 or control (IgG) immunoprecipitations from nuclear extracts of hESCs following the co-IP1 or co-IP2 protocols. Input is 5% of the material used for immunoprecipitations. Results are representative of two independent experiments. For gel Source Data, see Supplementary Fig. 1. **b**, Scatter plots of the  $\log_2$  ratios of label-free quantification (LFQ) intensities for proteins identified by quantitative mass spectrometry in SMAD2/3 co-immunoprecipitations compared with IgG negative control co-immunoprecipitations. The experiments were performed from nuclear extracts of hESCs. The SMAD2/3 and IgG negative control co-immunoprecipitations were differentially labelled after immunoprecipitation using the dimethyl method, followed by a combined run of the two samples in order to compare the abundance of specific peptides and identify enriched peptides. The values for technical dye-swap duplicates are plotted on different axes, and proteins whose enrichment was significant (significance  $B < 0.01$ ) are shown in black and named. As a result of this comparison between the two co-immunoprecipitation protocols, co-IP2 was selected for further experiments (see Supplementary

Discussion). **c**, Volcano plots of statistical significance against fold-change for proteins identified by label-free quantitative mass spectrometry in SMAD2/3 or IgG negative control immunoprecipitations in pluripotent hESCs or early endoderm (see Fig. 1a). The black lines indicate the threshold used to determine specific SMAD2/3 interactors, which are located to the right ( $n = 3$  co-immunoprecipitations; one-tailed  $t$ -test: permutation-based  $FDR < 0.05$ ). **d**, Selected results of the analysis described in **c** for SMAD2, SMAD3 and selected known bona fide SMAD2/3-binding partners (full results can be found in Supplementary Table 1). **e**, Mean label-free quantification (LFQ) intensity  $\log_2$  ratios in endoderm (endo) and pluripotent cells (pluri) for all SMAD2/3 interactors. Differentially enriched proteins are shown as green and blue bars. **f**, Selected results from gene ontology (GO) enrichment analysis, and enrichment analysis for mouse phenotypes annotated in the MGI database. All putative SMAD2/3-interacting proteins were considered for this analysis ( $n = 89$  proteins; Fisher's exact test followed by Benjamini–Hochberg correction for multiple comparisons). For each term, its rank in the analysis, the adjusted  $P$  value, and the number of associated genes are reported.

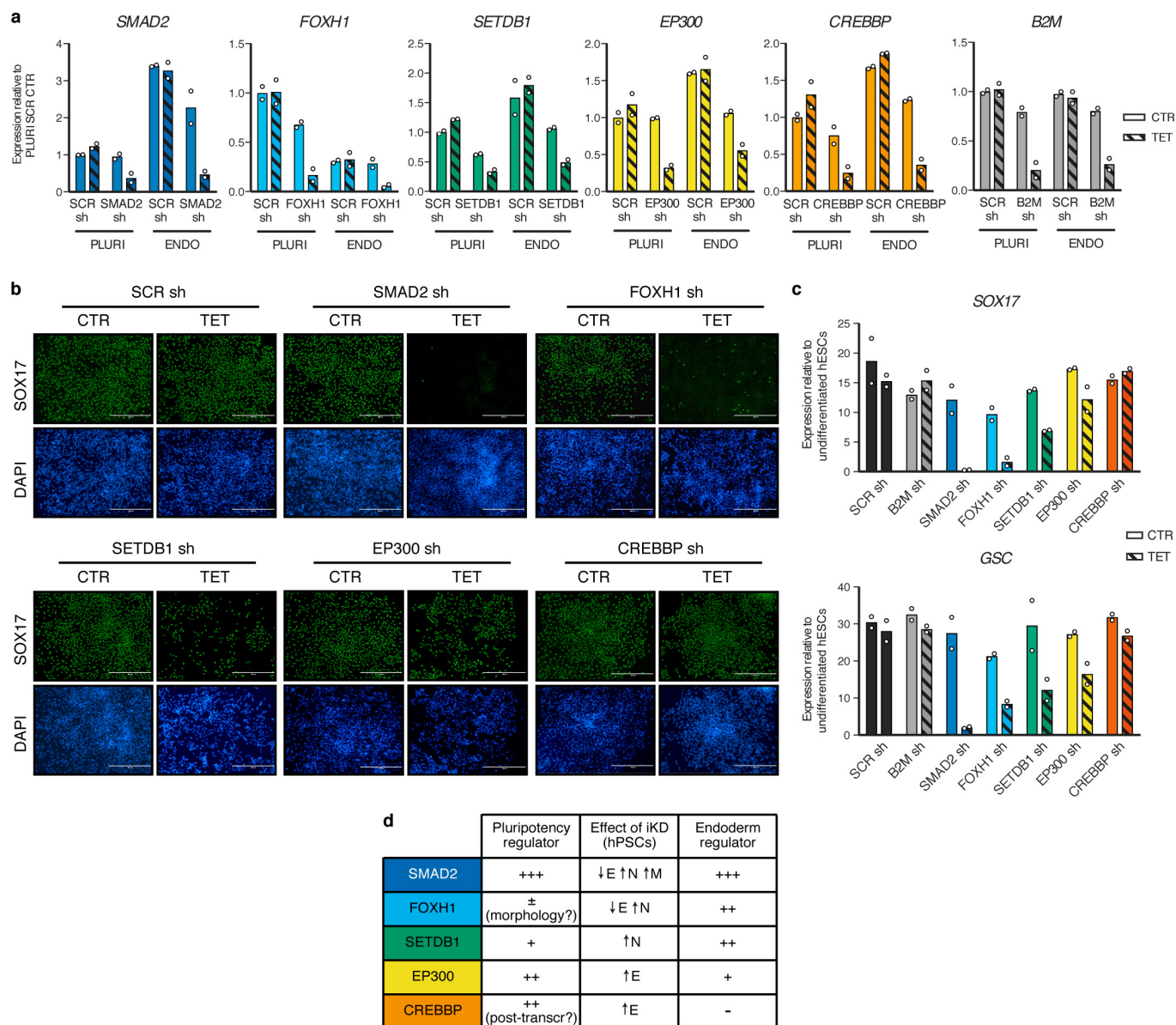


Extended Data Figure 2 | See next page for caption.

**Extended Data Figure 2 | Functional characterization of transcriptional and epigenetic cofactors of SMAD2/3 in hPSCs.** **a**, Western blots of SMAD2/3 or control (IgG) immunoprecipitations from nuclear extracts of pluripotent hESCs (pluri) or hESCs differentiated into endoderm for 36 h (endo). Input is 5% of the material used for immunoprecipitations. Results are representative of two independent experiments. **b**, Schematic of the experimental approach for the generation of iKD hESC lines for SMAD2/3 cofactors. **c**, qPCR screening of iKD hESCs cultured in the absence (CTR) or presence (TET) of tetracycline for three days. Three distinct shRNAs were tested for each gene. Expression is normalized to the mean level in hESCs carrying negative control shRNAs (scrambled (SCR) or against B2M) and cultured in the absence of tetracycline. The mean is indicated,  $n = 2$  independent clonal pools. Note that for the

B2M shRNA only the scrambled shRNA was used as negative control. shRNAs selected for further experiments are circled. **d**, Phase-contrast images of iKD hESCs expressing the indicated shRNAs (sh) and cultured in the presence of tetracycline for six days to induce knockdown. Scale bars, 400  $\mu\text{m}$ . Results are representative of two independent experiments. **e**, Immunofluorescence for the pluripotency factor NANOG in iKD hESCs for the indicated genes cultured in the absence (CTR) or presence of tetracycline (TET) for six days. Scale bars, 400  $\mu\text{m}$ . Results are representative of two independent experiments. **f**, Heat map summarizing qPCR analyses of iKD hESCs cultured as in **e**.  $\log_2$  fold-changes (FC) are compared to scrambled control ( $n = 2$  clonal pools). Germ-layer markers are grouped in boxes: green, endoderm; red, mesoderm; blue, neuroectoderm.



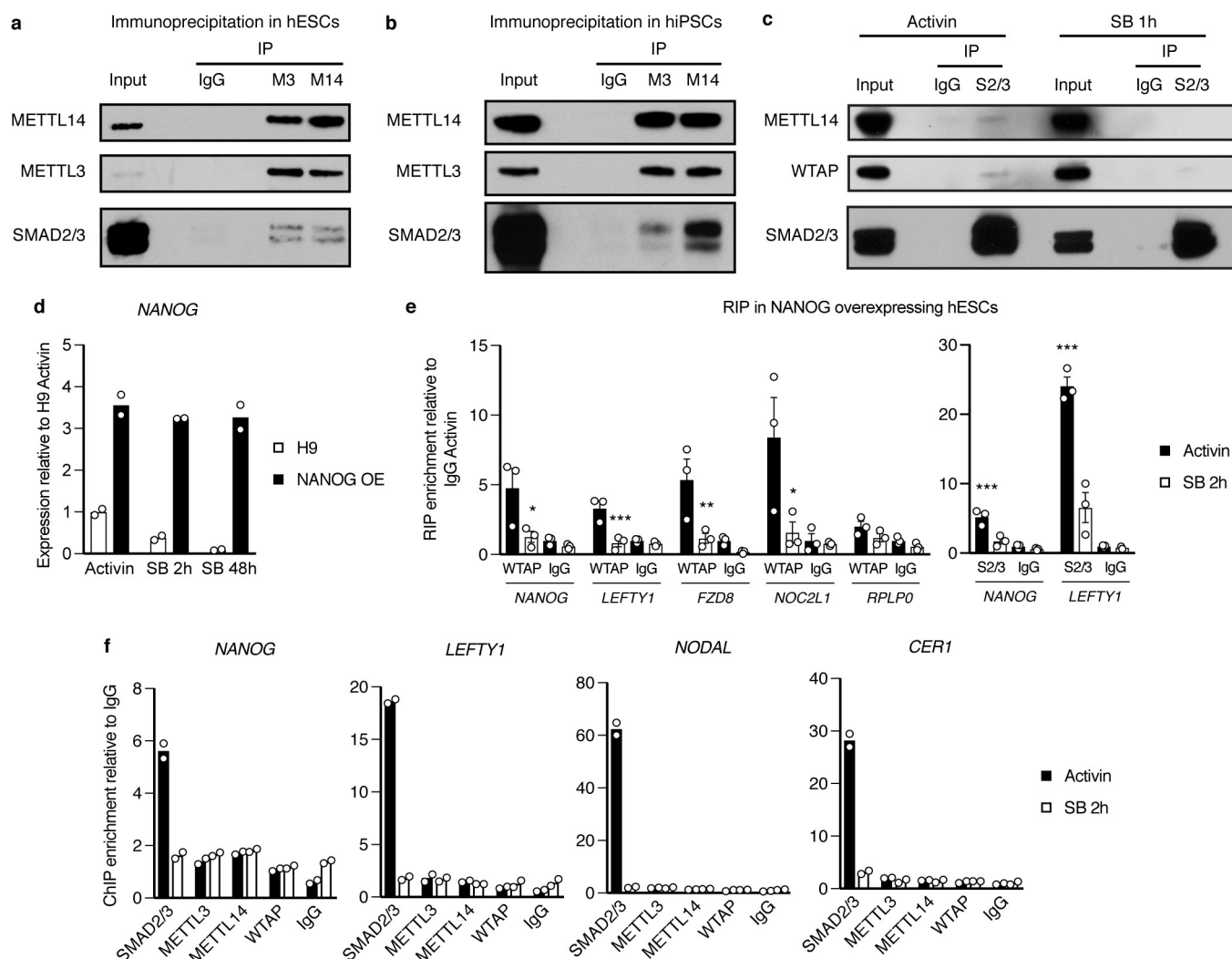


### Extended Data Figure 3 | Functional characterization of transcriptional and epigenetic cofactors of SMAD2/3 during endoderm differentiation.

**a**, qPCR validation of iKD hESCs in pluripotent cells (PLURI) or following endoderm differentiation (ENDO). Pluripotent cells were cultured in the absence (CTR) or presence (TET) of tetracycline for six days. For endoderm differentiation, tetracycline treatment was initiated in undifferentiated hESCs for three days in order to ensure gene knockdown at the start of endoderm specification, and was then maintained during differentiation (three days). For each gene, the shRNA resulting in the strongest level of knockdown in hPSCs was selected (refer to Extended Data Fig. 2). Expression is normalized to the mean level in pluripotent

hESCs carrying scrambled control shRNA and cultured in the absence of tetracycline. The mean is indicated,  $n = 2$  independent clonal pools.

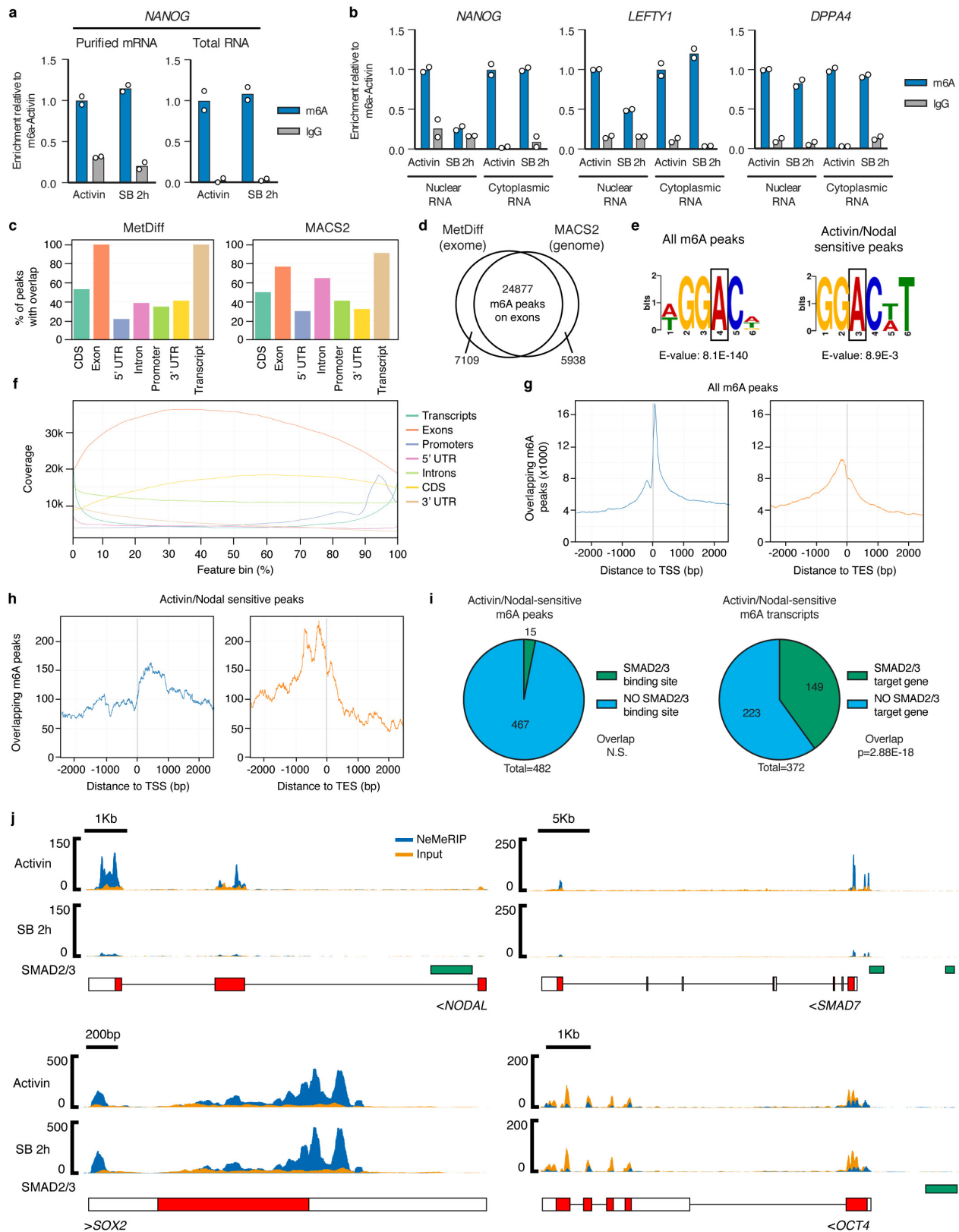
**b**, Immunofluorescence of the endoderm marker SOX17 following endoderm differentiation of iKD hESCs expressing the indicated shRNAs and cultured as described in **a**. Scale bars, 400  $\mu\text{m}$ . Results are representative of two independent experiments. **c**, qPCR following endoderm differentiation of iKD hESCs. The mean is indicated,  $n = 2$  independent clonal pools. **d**, Table summarizing the phenotypic results presented in Extended Data Fig. 2 and in this figure. E, endoderm; N, neuroectoderm; M, mesoderm.



#### Extended Data Figure 4 | Mechanistic insights into the functional interaction between SMAD2/3 and the m<sup>6</sup>A methyltransferase complex.

**a–c**, Western blots of SMAD2/3, METTL3, METTL14 or control (IgG) immunoprecipitations from nuclear extracts of hESCs (**a**, **c**) or hiPSCs (**b**). Input is 5% of the material used for immunoprecipitations. In **c**, immunoprecipitations were performed from hPSCs maintained in the presence of activin or treated for 1 h with the activin–NODAL signalling inhibitor SB431542. Results are representative of three (**a**) or two (**b**, **c**) independent experiments. **d**, qPCR validation of hESCs constitutively overexpressing NANOG (NANOG OE) following gene targeting of the *AAVS1* locus with pAAV-Puro\_CAG-NANOG. Parental wild-type H9 hESCs (H9) were analysed as negative controls. Cells were cultured in the presence of activin or treated with SB431542 for the indicated times. The mean is shown,  $n = 2$  cultures. NANOG-overexpressing cells are resistant to downregulation of NANOG following inhibition of activin–

NODAL signalling. **e**, RNA immunoprecipitation experiments for WTAP, SMAD2/3 or IgG control in NANOG-overexpressing hESCs maintained in the presence of activin or treated for 2 h with SB431542. Enrichment of the indicated transcripts was measured by qPCR and expressed relative to background levels observed in control IgG RNA immunoprecipitations in the presence of activin. *RPLP0* was tested as a negative control transcript. Mean  $\pm$  s.e.m.,  $n = 3$  cultures. Significance of differences from activin (left) or IgG (right) RIP was tested by two-way ANOVA with post hoc Holm–Sidak comparisons. **f**, ChIP–qPCR in hESCs for ChIP against the indicated proteins or the negative control ChIP (IgG). qPCR was performed for validated genomic SMAD2/3-binding sites associated with the indicated genes<sup>10,30</sup>. hESCs were cultured in the presence of activin or treated for 2 h with SB431542. Enrichment is normalized against background binding observed with IgG ChIP. The mean is shown,  $n = 2$  technical replicates. Results are representative of three independent experiments.



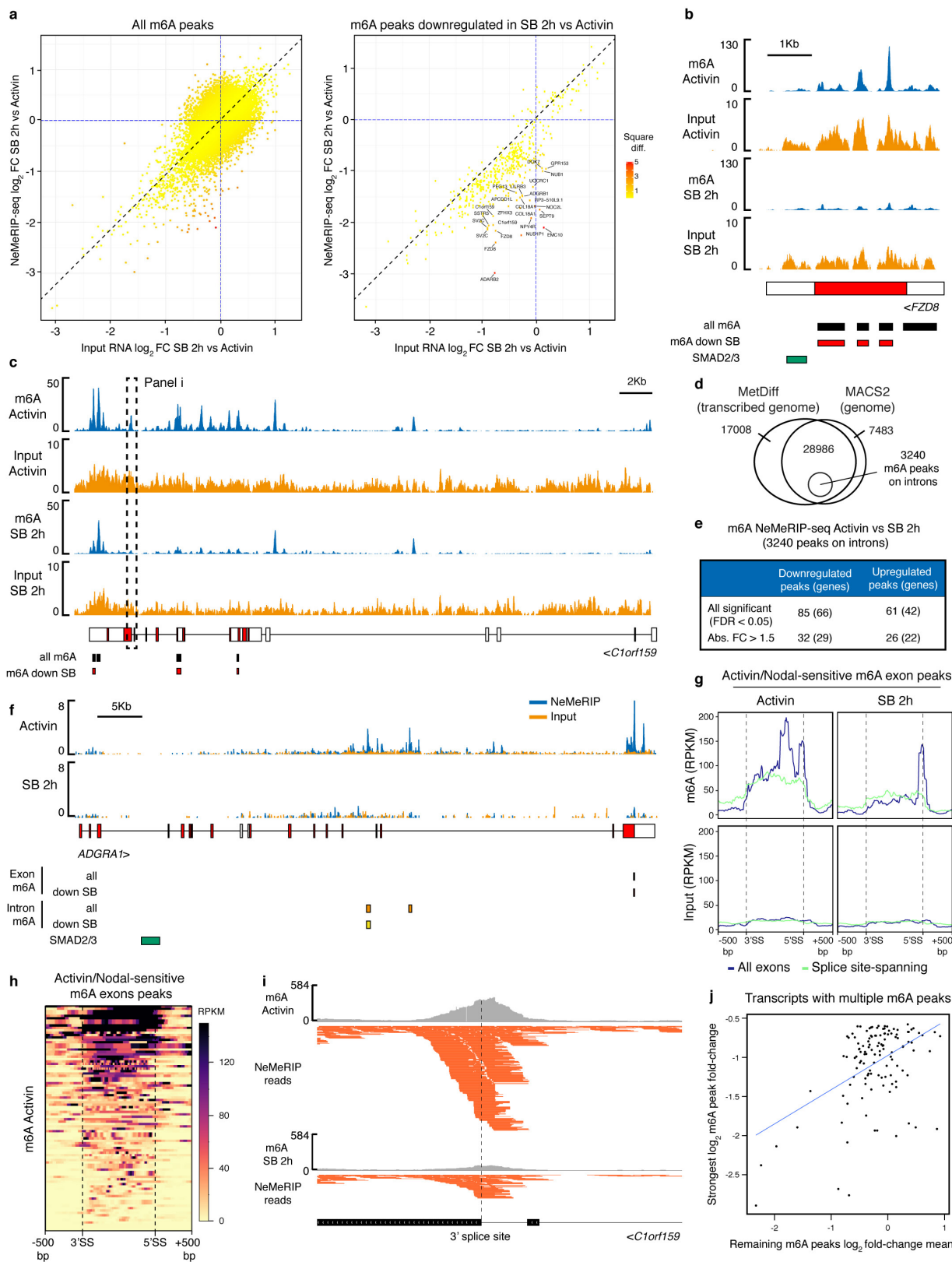
Extended Data Figure 5 | See next page for caption.

### Extended Data Figure 5 | Monitoring changes in m<sup>6</sup>A deposition that are rapidly induced by inhibition of activin–NODAL signalling.

**a, b**, MeRIP–qPCR results from purified mRNA, total cellular RNA or cellular RNA species separated by nuclear and cytoplasmic subcellular fractionation. hESCs were cultured in pluripotency-maintaining conditions containing activin or in conditions in which activin–NODAL signalling was inhibited for 2 h with SB431542. IgG MeRIP experiments were performed as negative controls. The mean is indicated,  $n = 2$  technical replicates. Differences between activin- and SB431542-treated cells were observed only in the nuclear-enriched fraction. Therefore, the nuclear-enriched MeRIP protocol (NeMeRIP) was used for subsequent experiments (refer to the Supplementary Discussion). Results are representative of two independent experiments. **c**, Overlap with the indicated genomic features of m<sup>6</sup>A peaks identified by NeMeRIP–seq using two different bioinformatics pipelines in which peak calling was performed using MetDiff or MACS2. For each pipeline, the analyses were performed on the union of peaks identified from data obtained in hESCs cultured in the presence of activin or with inhibition of activin–NODAL signalling for 2 h with SB431542 ( $n = 3$  cultures). Note that the sum of the percentages within each graph is not 100% because some m<sup>6</sup>A peaks overlap several feature types. MetDiff is an exome peak caller, and, accordingly, 100% of peaks map to exons. MACS2 identifies peaks throughout the genome. **d**, Venn diagrams showing the overlap of peaks identified by the two pipelines. Only MetDiff peaks that were also identified by MACS2 were considered for subsequent analyses focused on m<sup>6</sup>A peaks on exons. **e**, Top sequence motifs identified *de novo* on all m<sup>6</sup>A exon peaks, or on those that showed significant downregulation following inhibition of activin–NODAL signalling (activin–NODAL-sensitive m<sup>6</sup>A

peaks; Supplementary Table 2). The position of the methylated adenosine is indicated by a box. **f**, Coverage profiles for all m<sup>6</sup>A exon peaks across the length of different genomic features. Each feature type is expressed as 100 bins of equal length with 5' to 3' directionality. **g, h**, Overlap of m<sup>6</sup>A exon peaks and transcription start sites (TSS) or transcription end sites (TES). In **g**, the analysis was performed for all m<sup>6</sup>A peaks. In **h**, only activin–NODAL-sensitive peaks were considered. **i**, Left, activin–NODAL-sensitive m<sup>6</sup>A exon peaks were evaluated for direct overlap with SMAD2/3-binding sites as indicated by ChIP–seq<sup>30</sup>.  $n = 482$  peaks; FDR = 0.41 as calculated by the permutation test implemented by the GAT python package; N.S., not significant based on 95% confidence interval. Right, overlap was calculated after the same features were mapped to their corresponding transcripts or genes, respectively.  $n = 372$  genes; hypergeometric test  $P = 2.88 \times 10^{-18}$ , significant based on 95% confidence interval. **j**, m<sup>6</sup>A NeMeRIP–seq results for selected transcripts ( $n = 3$  cultures; replicates combined for visualization). Coverage tracks represent read enrichments normalized by million mapped reads and size of the library. Blue, sequencing results of m<sup>6</sup>A NeMeRIP; orange, sequencing results of pre-NeMeRIP input RNA (negative control). GENCODE gene annotations are shown (red, protein coding exons; white, untranslated exons; all potential exons are shown and overlaid). The location of SMAD2/3 ChIP–seq-binding sites is also shown. Compared to the other genes shown, m<sup>6</sup>A levels on *SOX2* were unaffected by inhibition of activin–NODAL signalling, showing specificity of action. *POU5F1* (also known as *OCT4*) is used as a negative control since it is known to not have a m<sup>6</sup>A site<sup>23</sup>, as confirmed by the lack of m<sup>6</sup>A enrichment compared to the input.

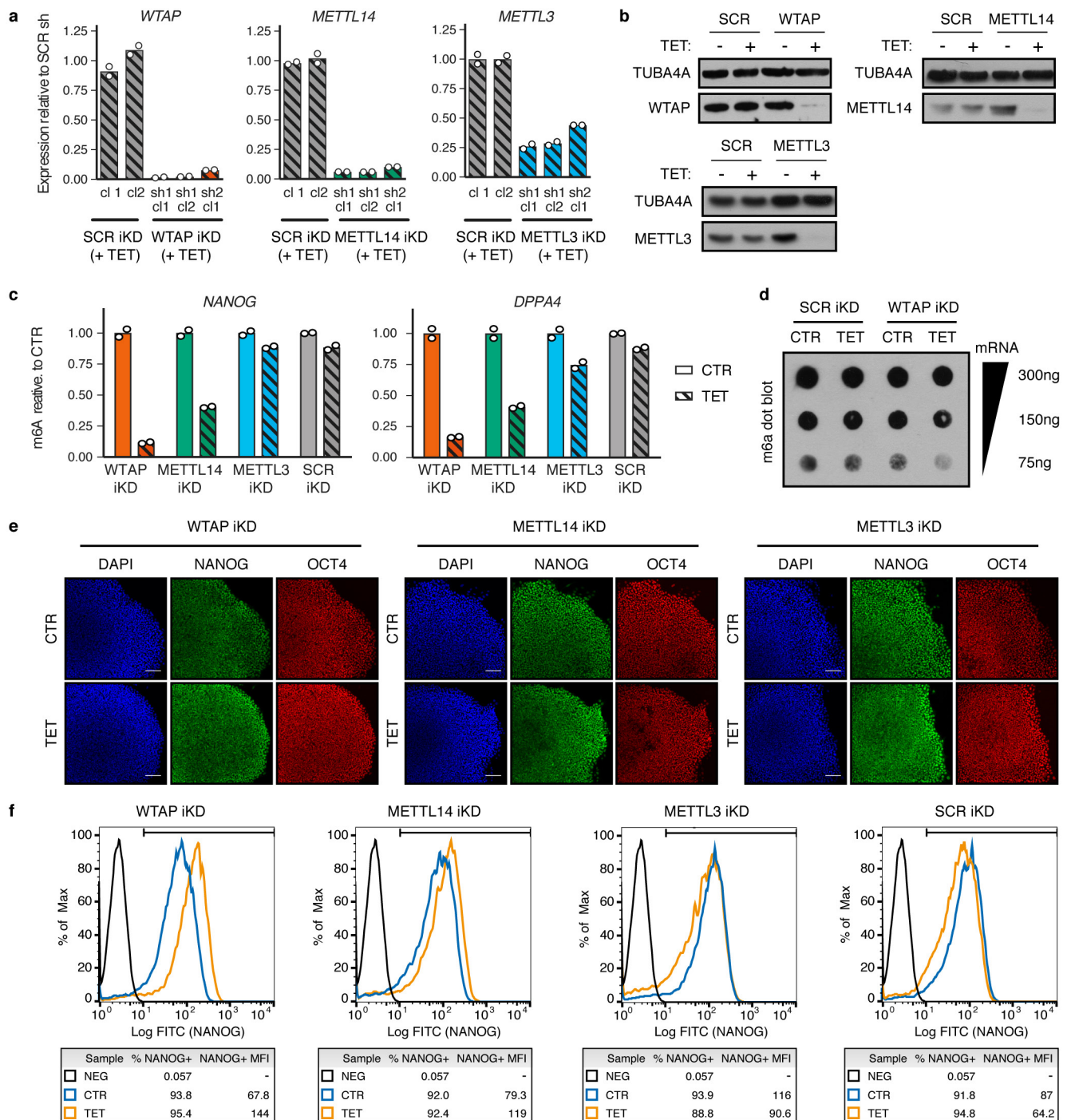




Extended Data Figure 6 | See next page for caption.

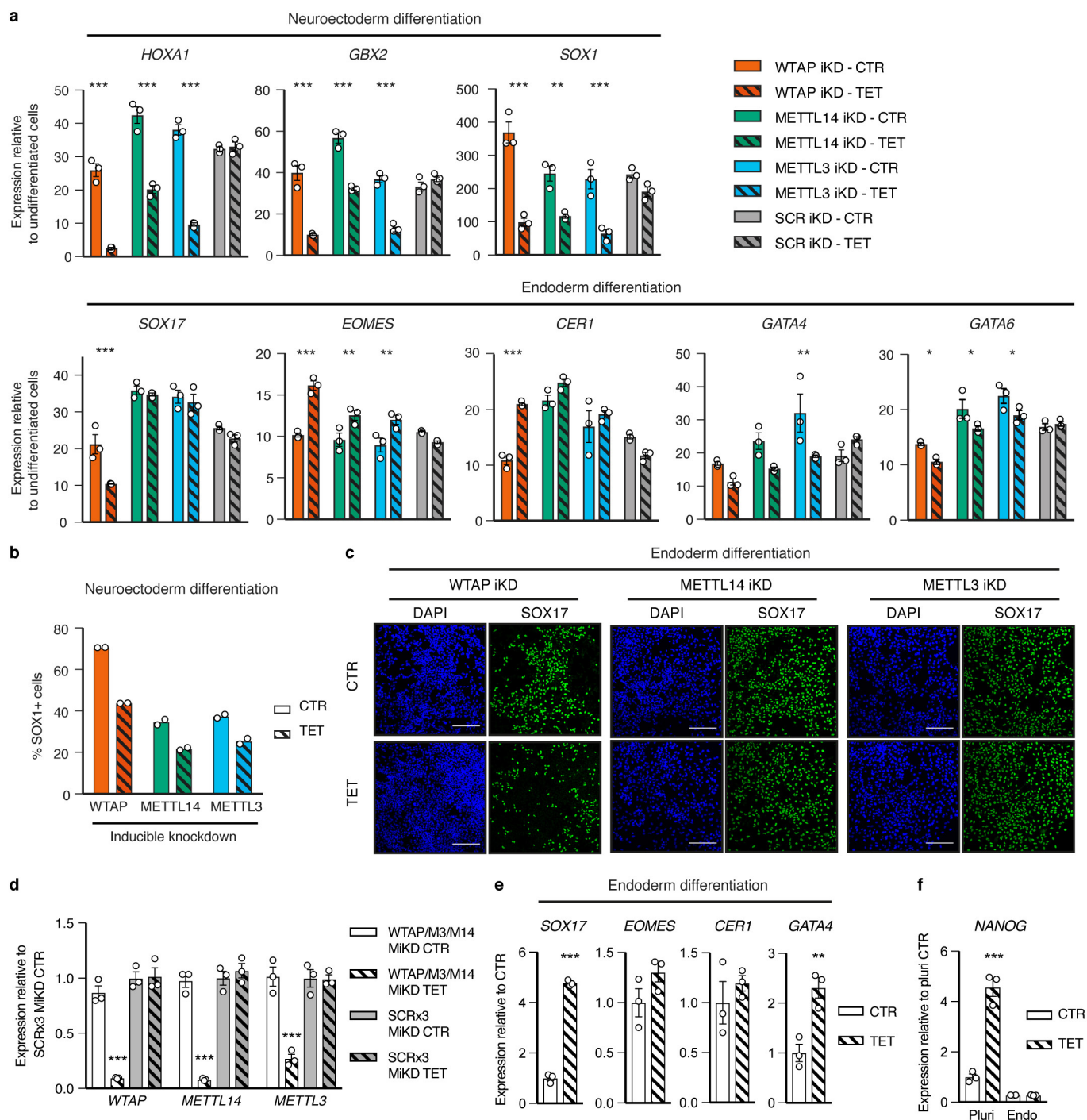
**Extended Data Figure 6 | Features of activin–NODAL-sensitive differential m<sup>6</sup>A deposition.** **a**, Scatter plot of the average log<sub>2</sub> fold-change in SB431542 versus activin-treated hESCs for m<sup>6</sup>A NeMeRIP-seq and pre-NeMeRIP input RNA ( $n = 3$  cultures). The analysis was performed for all m<sup>6</sup>A exon peaks (left), or for those peaks that were significantly downregulated following inhibition of activin–NODAL signalling (right). Data are colour coded according to the square of the difference between the two values (square diff.). **b, c**, As Extended Data Fig. 5j, but for representative transcripts that are stably expressed following inhibition of activin–NODAL signalling for 2 h ( $n = 3$  cultures; replicates combined for visualization). The m<sup>6</sup>A NeMeRIP and input tracks were separated and are shown at different scales to facilitate comparison between the conditions. The m<sup>6</sup>A peaks and those significantly downregulated after 2 h of SB431542 treatment are indicated. **d**, Venn diagram illustrating the strategy for identification of m<sup>6</sup>A peaks on introns. Peaks mapping to the transcribed genome were obtained by running MetDiff using an extended transcriptome annotation based on the pre-NeMeRIP input RNA, which has a high abundance of introns. The resulting peaks were first filtered by overlap with genome-wide MACS2-identified peaks, and then by lack of overlap with annotated exons. **e**, Results of MetDiff differential methylation analysis in activin versus 2 h SB431542 treatment for m<sup>6</sup>A peaks on introns.  $n = 3$  cultures;  $P$  value calculated by likelihood ratio test implemented in the MetDiff R package, and adjusted to FDR by Benjamini–Hochberg correction. See Supplementary Table 2 for the FDR of individual peaks. Abs.FC, absolute fold-change. **f**, As Extended Data Fig. 5j, but for a representative transcript that shows activin–NODAL-sensitive m<sup>6</sup>A deposition in introns ( $n = 3$  cultures; replicates combined for visualization). The m<sup>6</sup>A peaks on exons, introns, and those

significantly downregulated after SB431542 treatment within each subset are indicated. **g**, Plots of RPKM-normalized mean m<sup>6</sup>A coverage for m<sup>6</sup>A exon peaks significantly downregulated after SB431542 treatment (absolute fold-change > 1.5). Data for all such peaks is in blue, whereas green lines report coverage for only those peaks characterized by next generation sequencing reads that span exon–intron junctions. Exons were scaled proportionally, and the positions of the 3' and 5' splice sites (SS) are indicated. A window of 500 bp on either side of the splice sites is shown. m<sup>6</sup>A, signal from m<sup>6</sup>A NeMeRIP-seq; input, signal from pre-NeMeRIP input RNA. The results show that coverage of activin–NODAL-sensitive m<sup>6</sup>A peaks often spans across splice sites (highlighted by the dotted lines). **h**, Heat map representing in an extended form the data shown in **g** for all activin–NODAL-sensitive m<sup>6</sup>A exon peaks in hESCs cultured in the presence of activin. There are multiple regions in which sequencing coverage extends across exon–intron junctions (see Supplementary Table 2). **i**, Example of an activin–NODAL-sensitive peak located in the proximity of a 3' splice site ( $n = 3$  cultures; replicates combined for visualization). This peak is shown within its genomic context in **c**, where it is indicated by a dotted box. Top, m<sup>6</sup>A NeMeRIP-seq coverage; bottom, individual next generation sequencing reads. Multiple reads span the exon–intron junction (indicated by the dashed line). **j**, Relationship between the decrease of m<sup>6</sup>A on the most affected exonic peak located on a transcript ( $y$  axis) and the mean change of all other peaks mapping to the same transcript ( $x$  axis). The analysis considered transcripts with multiple m<sup>6</sup>A peaks and with at least one peak significantly decreasing after inhibition of activin–NODAL signalling with SB431542 (absolute fold-change > 1.5). Sensitivity of m<sup>6</sup>A deposition to activin–NODAL signalling across these transcripts is correlated.



**Extended Data Figure 7 | Generation and functional characterization of hPSCs following iKD of the subunits of the m<sup>6</sup>A methyltransferase complex.** **a**, qPCR validation of iKD hESCs cultured in the presence of tetracycline for five days (TET) to drive gene knockdown. Two distinct shRNAs and multiple clonal sublines (cl) were tested for each gene. Expression is normalized to the mean level in hESCs carrying a negative control scrambled (SCR) shRNA. For each gene, sh1 cl1 was selected for further analyses. The mean is indicated,  $n = 2$  cultures. **b**, Western blot validation of selected iKD hESCs for the indicated genes. TUBA4A ( $\alpha$ -tubulin), loading control. Results are representative of three independent experiments. **c**, MeRIP-qPCR in iKD hESCs cultured for ten days in the absence (CTR) or presence of tetracycline (TET). m<sup>6</sup>A abundance is shown relative to control conditions in the same hESC line. The mean is shown,  $n = 2$  technical replicates. Results are representative

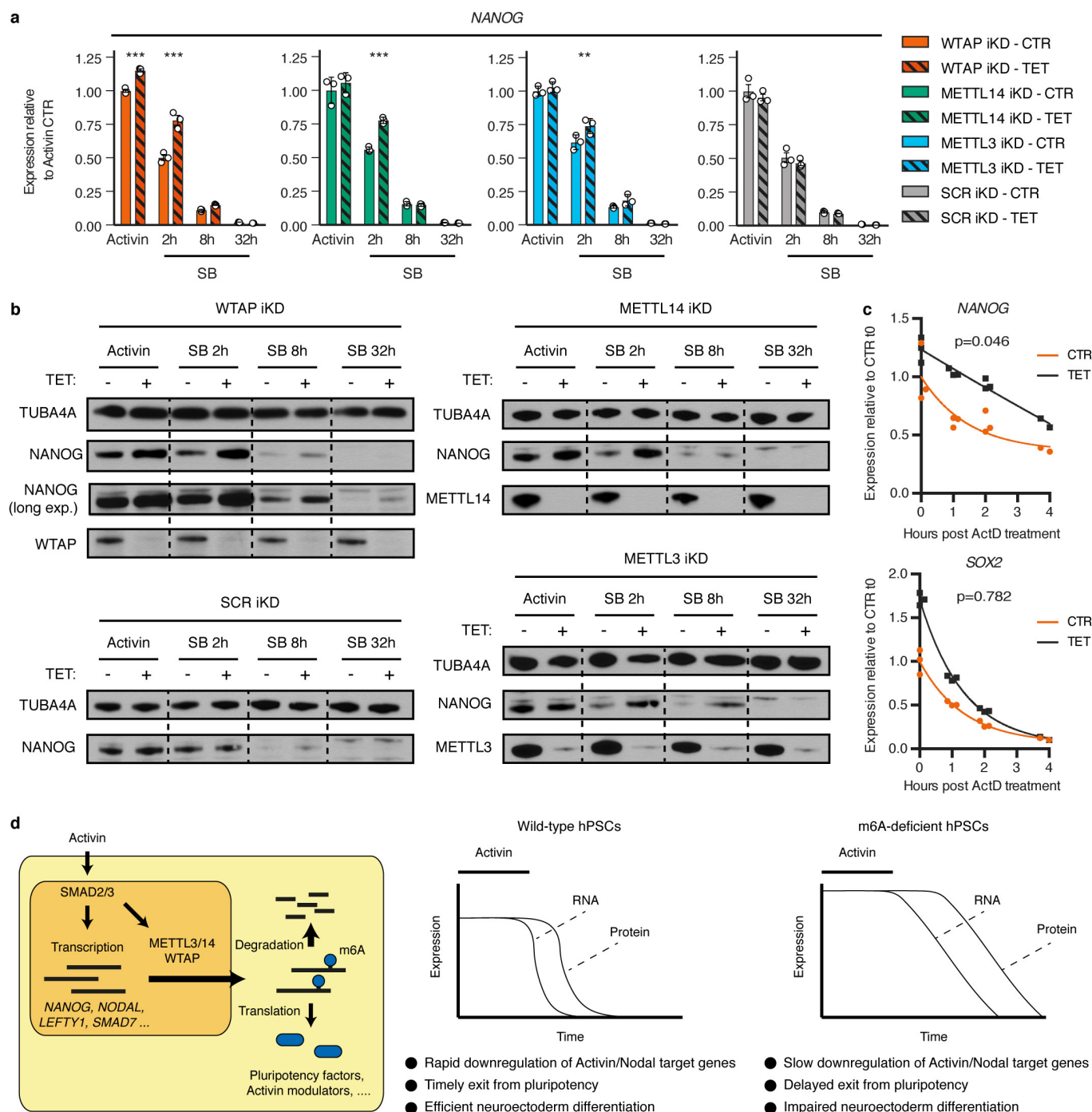
of two independent experiments. **d**, m<sup>6</sup>A dot blot in WTAP or scramble shRNA control iKD hESCs treated as described in **c**. Decreasing amounts of mRNA were spotted to allow semiquantitative comparisons, as indicated. Results are representative of two independent experiments. **e**, Immunofluorescence of the pluripotency markers NANOG and OCT4 in iKD hESCs cultured for three passages (15 days) in the absence (CTR) or presence of tetracycline (TET). Scale bars, 100  $\mu$ m. Results are representative of two independent experiments. **f**, Flow cytometry showing NANOG expression in cells treated as in **e**. The percentage and median fluorescence intensity (MFI) of NANOG-positive cells (NANOG<sup>+</sup>) are shown. The gates used for the analysis are indicated, and were determined on the basis of a secondary-antibody-only negative staining (NEG). Results are representative of two independent experiments.



**Extended Data Figure 8 | Function of the m<sup>6</sup>A methyltransferase complex during germ-layer specification.** **a**, qPCR analysis following neuroectoderm or endoderm differentiation of iKD hESCs cultured in absence (CTR) or presence of tetracycline (TET). Tetracycline treatment was initiated in undifferentiated hESCs for ten days and was maintained during differentiation (three days). Expression was normalized against the mean level in undifferentiated hESCs. Mean  $\pm$  s.e.m.,  $n = 3$  cultures. Significant differences versus the same iKD line in control conditions were calculated by two-way ANOVA with post hoc Holm–Sidak comparisons. **b**, Flow cytometry quantification of the percentage of SOX1<sup>+</sup> cells (SOX1<sup>+</sup>) in cells treated as in **a**. Mean is shown,  $n = 2$  cultures. **c**, Immunofluorescence of the lineage marker SOX17 in endoderm-differentiated hESCs treated as in **a**. Scale bars, 100  $\mu$ m. Results are

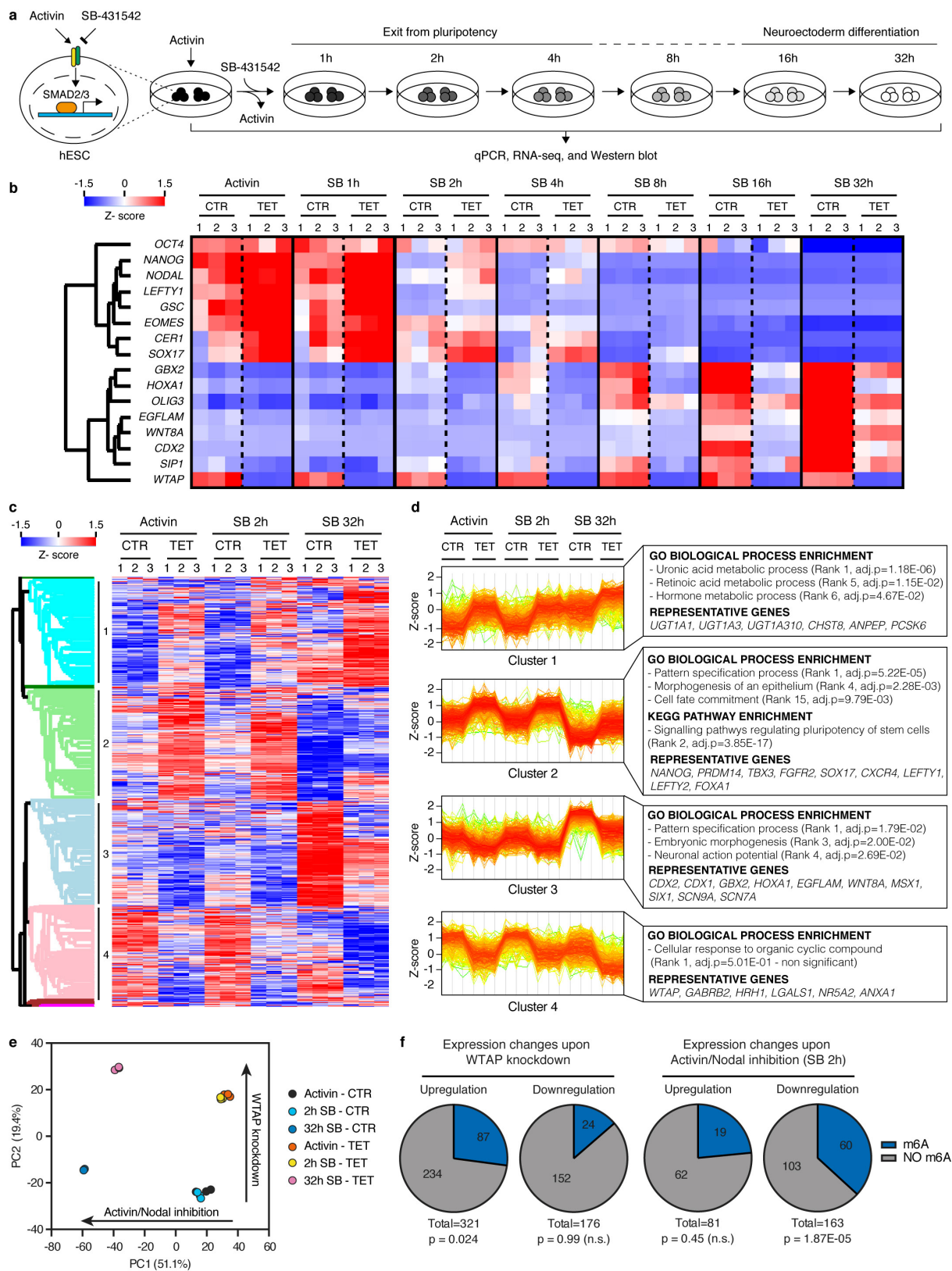
representative of two independent experiments. **d**, qPCR validation of multiple inducible knockdown (MiKD) hESCs simultaneously expressing shRNAs against WTAP, METTL3 and METTL14. Cells expressing three copies of the scrambled shRNA (SCR3 $\times$ ) were used as negative control. Cells were cultured in the presence of tetracycline (TET) for five days to drive gene knockdown. Mean  $\pm$  s.e.m.,  $n = 3$  cultures. Significant differences versus SCR3 $\times$  hESCs in control conditions were calculated by two-way ANOVA with post hoc Holm–Sidak comparisons. **e**, **f**, qPCR analysis following endoderm differentiation of WTAP, METTL3 and METTL14-MiKD hESCs treated as described in **a**. Mean  $\pm$  s.e.m.,  $n = 3$  cultures. Significant differences versus control conditions were calculated by two-tailed  $t$ -test (**e**) or two-way ANOVA with post hoc Holm–Sidak comparisons (**f**).





**Extended Data Figure 9 | Function of the m<sup>6</sup>A methyltransferase complex during exit from pluripotency induced by inhibition of activin–NODAL signalling.** **a**, qPCR analyses in iKD hESCs cultured in absence (CTR) or presence (TET) of tetracycline for ten days, then subjected to inhibition of activin–NODAL signalling with SB431542 (SB) for the indicated time (see Extended Data Fig. 10a). Activin, cells maintained in standard pluripotency-promoting culture conditions containing activin and collected at the beginning of the experiment. Mean  $\pm$  s.e.m.,  $n = 3$  cultures. Significant differences versus same iKD line in control conditions were calculated by two-way ANOVA with post hoc Holm–Sidak comparisons. **b**, Western blots of cells treated as described in **a**. TUBA4A, loading control. Results are representative of two independent

experiments. **c**, Measurement of mRNA stability in WTAP iKD hESCs cultured in absence (CTR) or presence (TET) of tetracycline for ten days. Samples were collected following transcriptional inhibition using actinomycin D (ActD) for the indicated time. The statistical significance of differences between the mRNA half lives in tetracycline versus control is shown ( $n = 3$  cultures, comparison of fits to one-phase decay model by extra sum-of-squares  $F$ -test). The difference was significant for *NANOG* but not for *SOX2* (95% confidence interval). **d**, Model showing the interplays between activin–NODAL signalling and m<sup>6</sup>A deposition in hPSCs (left), and the phenotype induced by impairment of the m<sup>6</sup>A methyltransferase complex (right).



Extended Data Figure 10 | See next page for caption.

**Extended Data Figure 10 | Genome-wide analysis of the relationship between WTAP and activin–NODAL signalling.** **a**, Schematic of the experimental approach to investigate the transcriptional changes induced by the knockdown of the m<sup>6</sup>A methyltransferase complex subunits during neuroectoderm specification of hESCs. **b**, qPCR analyses of WTAP iKD hESCs subjected to the experiment outlined in **a** ( $n = 3$  cultures). Activin, cells maintained in standard pluripotency-promoting culture conditions containing activin and collected at the beginning of the experiment. Z-scores indicate differential expression measured in number of standard deviations from the mean across all time points. **c**, RNA-seq analysis at selected time points from the samples shown in panel **b** ( $n = 3$  cultures). The heat map shows Z-scores for the top 5% differentially expressed genes (1789 genes as ranked by the Hotelling  $T^2$  statistic). Genes and samples were clustered based on their Euclidean distance, and the four major gene clusters are indicated (see Supplementary Discussion). **d**, Expression profiles of genes belonging to the clusters indicated in **c**. Selected results of gene-enrichment analysis and representative genes for each cluster

are shown (cluster 1:  $n = 456$  genes; cluster 2:  $n = 471$  genes; cluster 3:  $n = 442$  genes; cluster 4:  $n = 392$  genes; Fisher's exact test followed by Benjamini–Hochberg correction for multiple comparisons). **e**, Principal component analysis of RNA-seq results in **c** ( $n = 3$  cultures). The top 5% differentially expressed genes were considered for this analysis. For each of the two main principal components (PC1 and PC2), the fraction of inter-sample variance that they explain and their proposed biological meaning are reported. **f**, Proportion of transcripts marked by at least one high-confidence m<sup>6</sup>A peak<sup>23</sup> in transcripts significantly up- or downregulated following WTAP iKD in hESCs maintained in the presence of activin (left), or following inhibition of activin–NODAL signalling for 2 h with SB431542 in control cells (right). Differential gene expression was calculated in three cultures using the negative binomial test implemented in DESeq2 with a cutoff of  $P < 0.05$  and  $\text{abs.FC} > 2$ . The number of genes in each group and the hypergeometric probabilities of the observed overlaps with m<sup>6</sup>A-marked transcripts are reported (n.s.: not significant at 95% confidence interval).

# Hierarchical roles of mitochondrial Papi and Zucchini in *Bombyx* germline piRNA biogenesis

Kazumichi M. Nishida<sup>1\*</sup>, Kazuhiro Sakakibara<sup>1\*</sup>, Yuka W. Iwasaki<sup>2</sup>, Hiromi Yamada<sup>1</sup>, Ryo Murakami<sup>1</sup>, Yukiko Murota<sup>1</sup>, Takeshi Kawamura<sup>3,4</sup>, Tatsuhiko Kodama<sup>4</sup>, Haruhiko Siomi<sup>2</sup> & Mikiko C. Siomi<sup>1</sup>

PIWI-interacting RNAs (piRNAs) are small regulatory RNAs that bind to PIWI proteins to control transposons and maintain genome integrity in animal germ lines<sup>1–3</sup>. piRNA 3' end formation in the silkworm *Bombyx mori* has been shown to be mediated by the 3'-to-5' exonuclease Trimmer (Trim; known as PNLDC1 in mammals)<sup>4</sup>, and piRNA intermediates are bound with PIWI anchored onto mitochondrial Tudor domain protein Papi<sup>5</sup>. However, it remains unclear whether the Zucchini (Zuc) endonuclease and Nibbler (Nbr) 3'-to-5' exonuclease, both of which have pivotal roles in piRNA biogenesis in *Drosophila*<sup>6–8</sup>, are required for piRNA processing in other species. Here we show that the loss of Zuc in *Bombyx* had no effect on the levels of Trim and Nbr, but resulted in the aberrant accumulation of piRNA intermediates within the Papi complex, and that these were processed to form mature piRNAs by recombinant Zuc. Papi exerted its RNA-binding activity only when bound with PIWI and phosphorylated, suggesting that complex assembly involves a hierarchical process. Both the 5' and 3' ends of piRNA intermediates within the Papi complex showed hallmarks of PIWI 'slicer' activity, yet no phasing pattern was observed in mature piRNAs. The loss of Zuc did not affect the 5'- and 3'-end formation of the intermediates, strongly supporting the idea that the 5' end of *Bombyx* piRNA is formed by PIWI slicer activity, but independently of Zuc, whereas the 3' end is formed by the Zuc endonuclease. The *Bombyx* piRNA biogenesis machinery is simpler than that of *Drosophila*, because *Bombyx* has no transcriptional silencing machinery that relies on phased piRNAs.

piRNAs are produced through an intricate biogenesis pathway composed of the primary pathway, the amplification loop (also known as the ping-pong cycle), and Zuc-dependent phasing<sup>1–3,6–12</sup>. To understand the mechanism that underlies the amplification machinery, the silkworm ovary-derived, cultured cell line BmN4 has been used<sup>13–15</sup>. BmN4 cells express two PIWI proteins, Siwi and Ago3. Siwi- and Ago3-bound piRNAs show strong nucleotide and strand biases (uracil at position 1 (1U) for the antisense strand, and adenine at position (10A) for the sense strand), and are complementary through 10 nucleotides from their 5' end, known as the ping-pong signatures<sup>14,15</sup>. Both Siwi and Ago3 are cytoplasmic: that is, silkworms rely on solely post-transcriptional silencing to control transposons, unlike *Drosophila* and mice, which repress transposons both transcriptionally and post-transcriptionally<sup>1–3</sup>.

Papi<sup>16</sup> in BmN4 cells is anchored to the surface of mitochondria through a mitochondrial localization signal (MLS), and binds Siwi and Ago3 through their symmetrically dimethylated arginine (sDMA) residues<sup>5</sup>. Depletion of Papi was shown to have little effect on the levels of piRNAs, although piRNAs became several bases longer exclusively at their 3' end<sup>5</sup>. Trim was identified as the 3'-to-5' exonuclease required for piRNA 3' processing in *Bombyx*<sup>4</sup>.

Zuc endonuclease is necessary for piRNA biogenesis, particularly for phased piRNA production in *Drosophila* and mice<sup>6–8,17–19</sup>. Nbr exonuclease functions in piRNA 3'-end formation in *Drosophila*<sup>8,20</sup>. The silkworm genome contains homologues of Zuc and Nbr genes (KAIObase; <http://sgp.dna.affrc.go.jp>), both of which are expressed in BmN4 cells (Extended Data Fig. 1a, b). However, loss of Trim and Nbr did not affect the levels of Flag-PIWI-loaded piRNAs, whereas loss of Papi nearly completely abolished the PIWI–piRNA association (Fig. 1a and Extended Data Fig. 1a, c). Papi is essential for piRNA production and formation of the piRNA-induced silencing complex (piRISC) in silkworm germline cells.

PIWI-loaded piRNAs in Trim-depleted cells seemed to be subtly upshifted on RNA gels (Fig. 1a). The mean sizes of Siwi- and Ago3-bound piRNAs were 28.2 and 27.6 nucleotides (in control cells), and 28.6 and 28.2 nucleotides (in Trim-knockdown cells), respectively (Fig. 1b); that is, Siwi- and Ago3-bound piRNAs produced with no Trim function were on average 0.4 and 0.6 nucleotides longer than those in control cells. piRNA sequencing confirmed these results (Extended Data Fig. 2a). The 1U/10A and strand biases were greatly maintained after Trim depletion (Extended Data Fig. 2b, c), suggesting that the changes in piRNA size seem to be attributed to changes at the 3' end. The population of piRNAs was also barely changed by Trim depletion (Extended Data Fig. 2c). In addition, lack of Nbr caused no notable changes in Siwi- and Ago3-bound piRNAs (Extended Data Fig. 2a–c). Trim may thus act to fine-tune piRNA size at the 3' end.

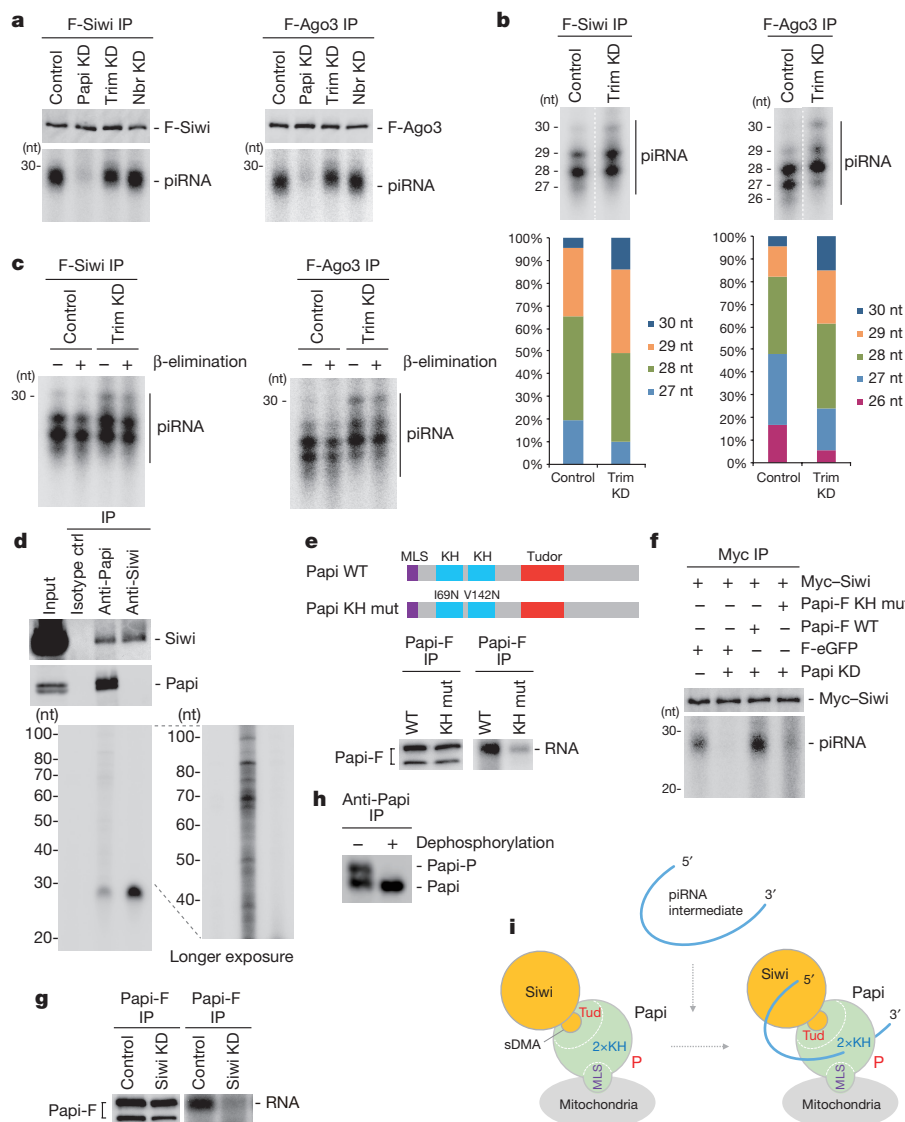
piRNAs undergo 2'-O-methylation by the methyltransferase Hen1 and become resistant to  $\beta$ -elimination<sup>21,22</sup>. Both Siwi- and Ago3-bound piRNAs, even longer ones, showed firm resistance to  $\beta$ -elimination, regardless of the presence or absence of Trim (Fig. 1c and Extended Data Fig. 3a). It seems that piRNA 2'-O-methylation occurs irrespective of Trim depletion. This is at odds with a previous report that claims that the lack of Trim impaired piRNA 3' end formation, including 2'-O-methylation, leading to a severe reduction in the piRNA level in silkworm cells<sup>4</sup>.

Mass spectrometric analysis of endogenous PIWI proteins in BmN4 cells revealed that 11 and 5 arginine residues of Siwi and Ago3, respectively, were sDMA-modified (Extended Data Fig. 3b, c and Supplementary Table 1). We substituted 9 out of the 11 arginine residues in Siwi to lysine residues (Siwi-9RK), which completely abolished the Siwi–Papi association (Extended Data Fig. 4a). The Ago3 mutant, which should have lost its sDMA modification, failed to bind Papi, as has been reported previously<sup>5</sup> (Extended Data Fig. 4a). Both mutants were barely loaded with piRNAs (Extended Data Fig. 4b), and failed to accumulate at nuage perinuclear foci<sup>23</sup>, the site for germline piRNA biogenesis (Extended Data Fig. 4c). Thus, sDMA modification is essential for the Papi association, nuage localization and piRISC formation of PIWI. Both Siwi and Ago3 were sDMA-modified even after

<sup>1</sup>Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0032, Japan. <sup>2</sup>Department of Molecular Biology, Keio University School of Medicine, Tokyo 162-8582, Japan. <sup>3</sup>Proteomics Laboratory, Isotope Science Center, The University of Tokyo, Tokyo 113-0032, Japan. <sup>4</sup>Laboratory for Systems Biology and Medicine, Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo 153-8904, Japan.

\*These authors contributed equally to this work.





**Figure 1 | Papi is essential for piRNA biogenesis and piRISC formation in Bmn4 cells.** **a**, Flag-tagged Siwi and Ago3 at the N termini (F-Siwi and F-Ago3, respectively) are loaded with piRNAs in Nbr-knockdown (KD) and Trim-knockdown but not Papi-knockdown Bmn4 cells. IP, immunoprecipitation. **b**, piRNAs appear to be slightly longer when Trim is depleted. **c**,  $\beta$ -elimination treatment of Siwi- and Ago3-bound piRNAs from Trim-knockdown and control cells. **d**, RNAs within the Papi complex and bound to Siwi were  $^{32}$ P-labelled at the 5' end. 'Isotype ctrl' denotes a non-immune IgG antibody. **e**, Wild-type (WT) Papi, but not the KH

mutant (mut) Papi, binds to RNAs. Papi-F, Flag-tagged Papi at the C terminus. **f**, Myc-Siwi is unloaded with piRNAs when the KH mutant is expressed in Papi-depleted cells. F-eGFP, Flag-tagged enhanced green fluorescent protein. **g**, Papi does not bind RNAs in Siwi-depleted Bmn4 cells. **h**, Papi is phosphorylated (Papi-P) in Bmn4 cells. **i**, Model showing that int-piRNA associates with the Papi-Siwi complex on mitochondria. Siwi-sDMA and Papi phosphorylation (P) are required for the assembly. Siwi may be replaced by Ago3 in this model. Tud, Tudor domain.

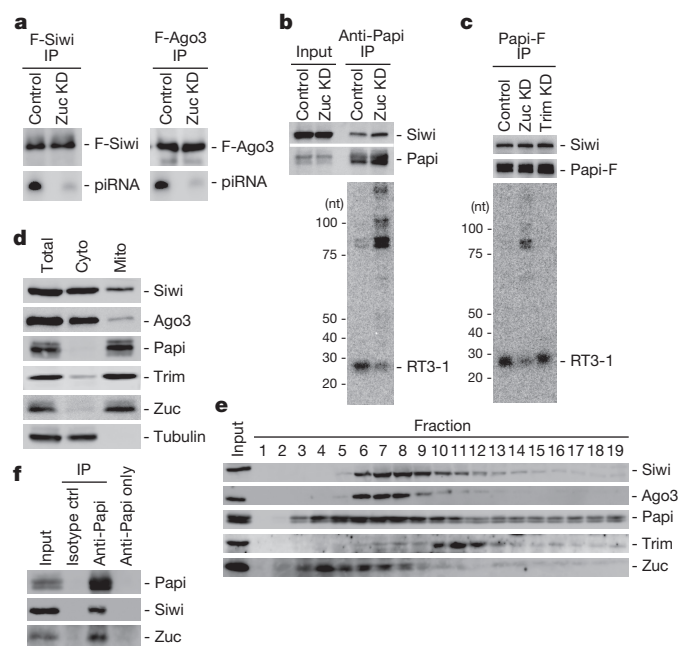
Papi depletion (Extended Data Fig. 4d), suggesting that the modification occurs before the Papi-PIWI association.

$^{32}$ P-labelling of RNAs within the Papi complex revealed that the levels of mature piRNAs in the complex were approximately 14.6% of those of piRNAs bound to Siwi itself (Fig. 1d), strongly supporting the idea that the piRISC is displaced from Papi after the completion of piRISC formation. The Papi complex contained not only Siwi but also Ago3 as expected, but the level of Ago3 was estimated to be about 10% of that of Siwi in the complex (Extended Data Fig. 5a).

The Papi complex contained RNA molecules longer than piRNAs (Fig. 1d), and this gave a positive result with a northern probe for RT3-1 (Extended Data Fig. 5b). RT3-1 was one of the abundant piRNAs loaded specifically onto Siwi<sup>14</sup>. piRNA intermediates (int-piRNAs) were detected similarly even after Siwi was forcibly displaced from Papi (Extended Data Fig. 5c). The int-piRNAs are therefore physically associated with Siwi in the Papi complex.

Papi contains two KH domains besides the Tudor domain and MLS (Fig. 1e). KH domains are RNA-binding motifs found in RNA-binding proteins that function in various types of RNA metabolism<sup>24</sup>. However, whether Papi exhibits its RNA-binding activity through KH domains has not been examined experimentally. Cross-linking immunoprecipitation (CLIP) experiments showed that Papi-Flag was efficiently cross-linked with RNAs in Bmn4 cells, as was endogenous Papi (Fig. 1e and Extended Data Fig. 5d). The Papi-Flag KH mutant, in which Ile69 and Val142 were mutated to asparagine residues, failed to bind RNAs (Fig. 1e). These two residues are highly conserved in KH domain-containing proteins and are crucial for the RNA-binding activity<sup>24</sup>. The Siwi-piRNA association was impeded when the Papi mutant was expressed instead of endogenous Papi (Fig. 1f and Extended Data Fig. 5e). Thus, the RNA-binding activity of Papi through KH domains is essential for piRISC formation.

Not only piRNAs but also int-piRNAs were hardly detected with the Siwi-9RK mutant (Extended Data Fig. 5f). Also, endogenous Siwi in



**Figure 2 | Zuc is essential for piRNA biogenesis and piRISC formation in BmN4 cells.** **a**, RNAs co-immunoprecipitated with Flag-Siwi and Flag-Ago3 are visualized by  $^{32}\text{P}$  labelling. **b**, RT3-1 int-piRNAs aberrantly accumulate in the Papi complex after Zuc depletion. **c**, Trim depletion does not affect the levels of RT3-1 int-piRNAs in the Papi complex. **d**, The presence of endogenous Siwi, Ago3, Papi, Trim and Zuc in the mitochondrial (mito) fraction. Cyto, cytoplasmic fraction. Tubulin was used as a loading control (bottom). **e**, The distribution patterns of endogenous Siwi, Ago3, Papi, Trim and Zuc in mitochondrial fractions 1–19 separated by sucrose gradient sedimentation. Fraction 1 contains the top (lightest) fraction. **f**, Zuc and Siwi are co-immunoprecipitated with Papi from mitochondrial lysates.

Papi-depleted cells was not loaded with either piRNAs or int-piRNAs (Extended Data Fig. 5g). Moreover, the intensity of the Papi CLIP signal decreased markedly when Siwi was depleted (Fig. 1g), although int-piRNAs were detected similarly in total RNAs irrespective of the presence of Siwi (Extended Data Fig. 5h). The Ago3–Papi association should be maintained after Siwi depletion. However, the CLIP signal was very low, agreeing with our earlier notion that the Ago3 level is low in the complex. It is likely that int-piRNAs join the Papi complex only after the Siwi–Papi association.

Both endogenous and exogenous Papi appeared as a doublet but only the top band was cross-linked with RNAs (Fig. 1e, g and Extended Data Fig. 5d). The top band does not represent a splicing variant because ectopically expressed Papi from the full-length cDNA appeared as a doublet. However, Papi became a single band after phosphatase treatment (Fig. 1h). Thus, Papi is subjected to phosphorylation and this modification is essential for it to bind RNAs.

The findings suggest that the assembly of the Papi–PIWI–int-piRNA complex occurs via a hierarchical process (Fig. 1i), which would possibly occur to ensure the funnelling of Siwi–piRNA intermediates to Papi–Siwi and Ago3–piRNA intermediates to the Papi–Ago3 complex. If Papi binds int-piRNA first, there must be a high chance that both Siwi and Ago3 would end up with the same set of piRNAs, disrupting the piRNA amplification. We hypothesized that the regulation of Papi phosphorylation is to keep Papi free from RNA until it associates with PIWI. However, this seems to be unlikely, given that Siwi depletion had a minimal effect on Papi phosphorylation (Fig. 1g).

Zuc depletion severely decreased the levels of Siwi- and Ago3-loaded piRNAs (Fig. 2a and Extended Data Fig. 6a). In sharp contrast, RT3-1 piRNA intermediates strongly accumulated in the Papi complex after Zuc depletion (Fig. 2b and Extended Data Fig. 5c). The depletion of Trim did not lead to the accumulation of int-piRNAs (Fig. 2c and

Extended Data Fig. 6b). The levels of Trim and *Nbr* (examined by western blotting and quantitative PCR, respectively) were unaffected by Zuc loss (Extended Data Fig. 6b, c). Thus, Zuc is responsible for piRNA processing in BmN4 cells.

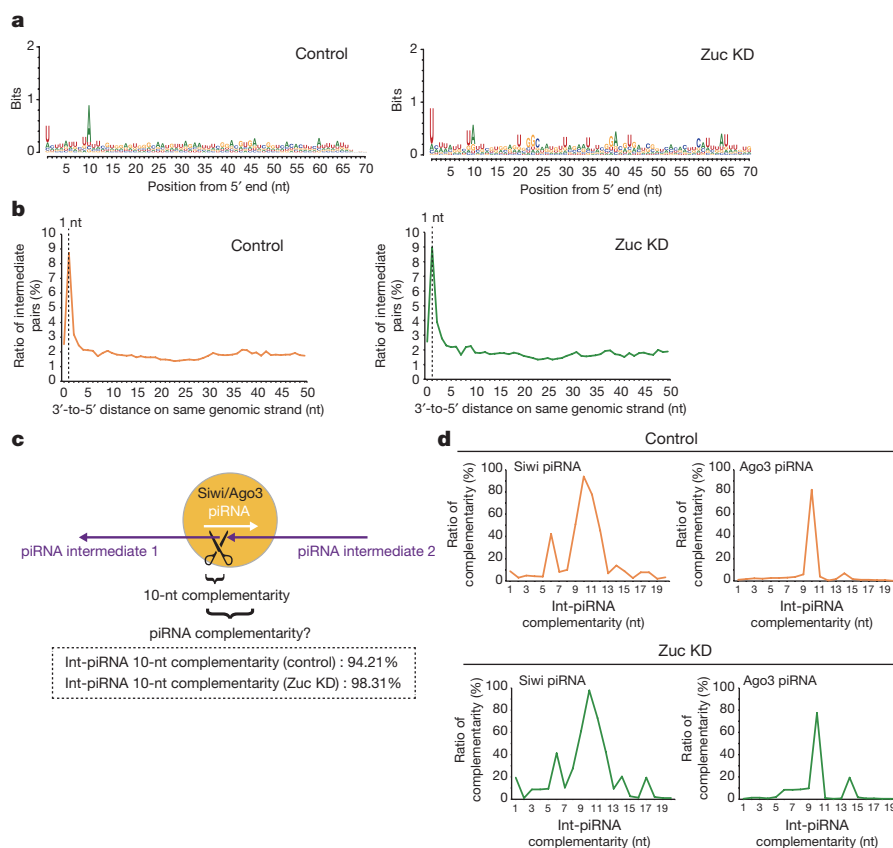
Papi, Zuc and Trim are present in the BmN4 mitochondrial fraction (Fig. 2d). In sucrose density-gradient experiments, Zuc and Papi were detected more strongly in lower-density fractions (Fig. 2e, fractions 3–8), whereas Trim was mostly found in higher-density ones (Fig. 2e, fractions 10–12). Zuc co-immunoprecipitated with Papi and Siwi (Fig. 2f). The mitochondrial Siwi complex contained Papi and Zuc (Extended Data Fig. 6d). The interaction of Zuc with the Papi–Siwi complex during the 3′-end piRNA processing was suggested.

We sequenced the libraries generated from 65–100-nucleotide-long RNAs extracted from the mitochondrial Papi complex before and after Zuc depletion (Extended Data Fig. 6e). More than 98% (98.09% in control and 98.65% in Zuc knockdown) of the reads contained one or more sequences of Siwi- or Ago3-bound piRNAs, indicating that the sequenced reads correspond to intermediate piRNAs. 1U/10A biases were detected in the intermediate piRNA reads (Fig. 3a), suggesting that piRNAs are generated from the 5′ end of the int-piRNA sequences. More than half of the intermediate piRNAs had piRNAs aligned to their 5′ end (51.94% for control and 59.58% for Zuc knockdown), suggesting that a large proportion of int-piRNAs produce piRNAs from their 5′ end. Apparent phasing pattern was not detected within piRNAs mapped to the same intermediate (Extended Data Fig. 6f), agreeing with the previous report demonstrating that exogenous piRNAs in BmN4 cells showed only weak phasing<sup>25</sup>. In *Drosophila*, phased piRNAs are loaded onto Piwi and transcriptionally control transposons. However, *Bombyx* lacks a Piwi homologue and relies solely on cytoplasmic PIWIs to silence transposons post-transcriptionally. Therefore, our finding that *Bombyx* produces no phased piRNAs appears reasonable.

We examined the distance from the 3′ end of each int-piRNA to the 5′ end of the next downstream intermediate, and found that the most common 3′-to-5′ distance was 1 nucleotide (Fig. 3b). This suggests that a single cleavage event, possibly by the Siwi or Ago3 slicer, produces the 3′ end of one int-piRNA and the 5′ end of the adjacent downstream int-piRNA, as in the case of the phased piRNAs. We then focused on the cleavage site of two adjacent int-piRNAs and analysed the population of piRNAs that possess 10-nucleotide complementarity at the cleavage site (Fig. 3c). More than 94% (94.21% for control and 98.31% for Zuc knockdown) of adjacent int-piRNA pairs had piRNAs that were complementary to 10 nucleotide from the 5′ end of the downstream int-piRNA and the 3′ end of the upstream int-piRNA. In addition, we calculated the proportion of int-piRNAs with complementary piRNA at each position from the 5′ end of the downstream int-piRNA, which was found to be highest at the 10-nt position for both Siwi- and Ago3-associated piRNAs (Fig. 3d). These results suggest that Siwi and Ago3 generate both 5′ and 3′ ends of int-piRNAs by slicer activity. This was observed in both control and Zuc-depleted BmN4 cells (Fig. 3), suggesting that Zuc is not involved in the production of both 5′ and 3′ ends of int-piRNAs.

Alteration of His169 at the active site of *Drosophila* Zuc to alanine abolished its endonuclease activity<sup>18</sup>. Incubation of recombinant wild-type *Bombyx* Zuc, but not the corresponding mutant His141Ala, with a naked 50-nucleotide RNA (1U50) produced 7–31-nucleotide RNAs dose-dependently (Fig. 4a and Extended Data Fig. 7a–c). Zuc may preferentially cleave after cytosine, and given that the 3′ side of guanine tends to be avoided, A–C/A and U–C/U/G may rarely be cleaved (Extended Data Fig. 7c). The cleavage pattern of an 80-nucleotide RNA, a 30-nucleotide extended version of 1U50 at the 3′ end, was very similar to that of 1U50 (Extended Data Fig. 7d, e). The predicted structures of the two RNA molecules were different (data not shown), suggesting that high-dimensional structures have a low effect on Zuc cleavage.

When RNA substrate was pre-loaded onto Flag-tagged Siwi, the product size was mostly in the range of 27–31 nucleotides (Fig. 4b and Extended Data Fig. 7f, g), which are typical or permissible sizes



**Figure 3 | Papi-associated int-piRNAs are generated by Siwi and Ago3 slicer.** **a**, Nucleotide bias of 65–100-nucleotide Papi-associated int-piRNAs in control and Zuc-depleted cells. **b**, Analyses of distance between Papi-associated intermediates in control and Zuc-depleted BmN4 cells. The distance between the 3' end of the upstream intermediate and the 5' end of the downstream intermediate on the same genomic strand is analysed. **c**, Illustration showing the cleavage of adjacent intermediates by Siwi or Ago3. The frequencies of the presence of a complementary piRNA

within 10 nucleotides from the 5' end of the downstream intermediate (piRNA intermediate 1) and 3' end of the upstream intermediate (piRNA intermediate 2) are indicated for control and Zuc-knockdown cells. **d**, Complementarity between piRNAs and adjacent intermediate sequence pairs. Graphs indicate the relative frequencies of the presence of a complementary piRNA (y axis), with the indicated distance at the downstream intermediate (x axis).

for silkworm piRNAs<sup>14</sup>. Similar results were obtained when the Siwi-RNA complex was pre-incubated with Papi (Fig. 4c and Extended Data Fig. 7h). Wild-type Zuc also processed RT3-1 int-piRNAs within the endogenous Papi complex to mature RT3-1 piRNA (Fig. 4d). Thus, Zuc endonuclease is the piRNA 3'-end processing factor in silkworm germline cells.

A new model for piRNA biogenesis in silkworm germline cells is shown in Extended Data Fig. 8a. In *Bombyx*, Papi might bind

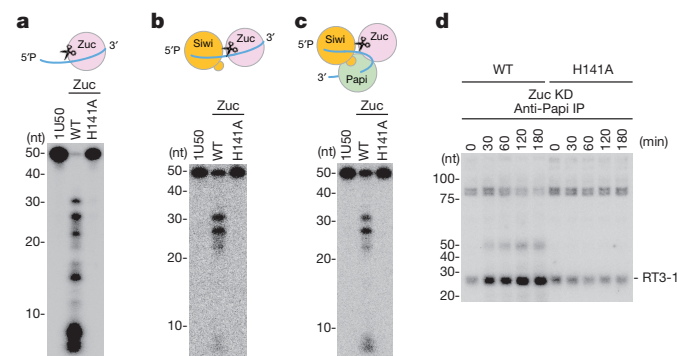
int-piRNAs towards the 3' end, whereas the int-piRNA 5' end is held by Siwi, most likely inserted into the 5' binding pocket<sup>26</sup> (Extended Data Fig. 8a, b). Under this structural arrangement, it would be nearly impossible for Trim or Nbr to process int-piRNAs from the very 3' end to mature piRNAs because Papi acts as an obstacle to the 3'-to-5' exonuclease reaction. Zuc is an endonuclease and so is able to process int-piRNAs even under such circumstances. Zuc shows only a subtle nucleotide preference in RNA cleavage. This unique trait helps the protein to act like an 'exonuclease', as a replacement of Nbr in *Drosophila*, to determine the length of piRNA in the biogenesis pathway.

In mouse testes, the loss of the Papi homologue TDRKH (also known as TDRD2) causes piRNA precursors to be accumulated on PIWI, because without Papi, nucleases responsible for piRNA 3'-end formation, such as Zuc (known as MITOPLD in mice) and/or Trim, are incapable of processing the 3' end<sup>17,19,27,28</sup>. In flies, the loss of Papi affected piRNA phasing, but hardly affected the levels of transposons in germ cells<sup>6,29</sup>. The piRNA pathway is highly conserved among animal species, but species diversity is evident from a mechanistic perspective.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 12 June 2017; accepted 26 January 2018.**

**Published online 28 February 2018.**



**Figure 4 | Zuc processes int-piRNAs to mature piRNAs.** **a**, Recombinant wild-type Zuc, but not the catalytically inactive Zuc(H141A) mutant, cleaves a naked 50-nucleotide RNA (1U50) *in vitro*. **b**, Wild-type Zuc cleaves 1U50 loaded onto Siwi. **c**, Wild-type Zuc cleaves 1U50 loaded onto the Papi-Siwi complex. **d**, Wild-type Zuc processes RT3-1 int-piRNAs accumulated in the Papi complex by Zuc depletion to mature RT3-1 piRNAs.

- Ghildiyal, M. & Zamore, P. D. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* **10**, 94–108 (2009).



2. Iwasaki, Y. W., Siomi, M. C. & Siomi, H. PIWI-interacting RNA: its biogenesis and functions. *Annu. Rev. Biochem.* **84**, 405–433 (2015).
3. Czech, B. & Hannon, G. J. One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. *Trends Biochem. Sci.* **41**, 324–337 (2016).
4. Izumi, N. *et al.* Identification and functional analysis of the pre-piRNA 3' Trimmer in silkworms. *Cell* **164**, 962–973 (2016).
5. Honda, S. *et al.* Mitochondrial protein BmPAPI modulates the length of mature piRNAs. *RNA* **19**, 1405–1418 (2013).
6. Han, B. W., Wang, W., Li, C., Weng, Z. & Zamore, P. D. Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. *Science* **348**, 817–821 (2015).
7. Mohn, F., Handler, D. & Brennecke, J. Noncoding RNA. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. *Science* **348**, 812–817 (2015).
8. Hayashi, R. *et al.* Genetic and mechanistic diversity of piRNA 3'-end formation. *Nature* **539**, 588–592 (2016).
9. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103 (2007).
10. Gunawardane, L. S. *et al.* A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science* **315**, 1587–1590 (2007).
11. Homolka, D. *et al.* PIWI slicing and RNA elements in precursors instruct directional primary piRNA biogenesis. *Cell Reports* **12**, 418–428 (2015).
12. Saito, K. *et al.* A regulatory circuit for *piwi* by the large Maf gene *traffic jam* in *Drosophila*. *Nature* **461**, 1296–1299 (2009).
13. Kawaoka, S. *et al.* The *Bombyx* ovary-derived cell line endogenously expresses PIWI/PIWI-interacting RNA complexes. *RNA* **15**, 1258–1264 (2009).
14. Nishida, K. M. *et al.* Respective functions of two distinct Siwi complexes assembled during PIWI-interacting RNA biogenesis in *Bombyx* germ cells. *Cell Reports* **10**, 193–203 (2015).
15. Xiol, J. *et al.* A role for Fkbp6 and the chaperone machinery in piRNA amplification and transposon silencing. *Mol. Cell* **47**, 970–979 (2012).
16. Liu, L., Qi, H., Wang, J. & Lin, H. PAPI, a novel TUDOR-domain protein, complexes with AGO3, ME31B and TRAL in the nuage to silence transposition. *Development* **138**, 1863–1873 (2011).
17. Ipsaro, J. J., Haase, A. D., Knott, S. R., Joshua-Tor, L. & Hannon, G. J. The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. *Nature* **491**, 279–283 (2012).
18. Nishimasu, H. *et al.* Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature* **491**, 284–287 (2012).
19. Watanabe, T. *et al.* MITOPLD is a mitochondrial protein essential for nuage formation and piRNA biogenesis in the mouse germline. *Dev. Cell* **20**, 364–375 (2011).
20. Feltzin, V. L. *et al.* The exonuclease Nibbler regulates age-associated traits and modulates piRNA length in *Drosophila*. *Aging Cell* **14**, 443–452 (2015).
21. Horwich, M. D. *et al.* The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr. Biol.* **17**, 1265–1272 (2007).
22. Saito, K. *et al.* Pimet, the *Drosophila* homolog of HEN1, mediates 2'-O-methylation of Piwi-interacting RNAs at their 3' ends. *Genes Dev.* **21**, 1603–1608 (2007).
23. Eddy, E. M. Germ plasm and the differentiation of the germ cell line. *Int. Rev. Cytol.* **43**, 229–280 (1975).
24. Nicastro, G., Taylor, I. A. & Ramos, A. KH-RNA interactions: back in the groove. *Curr. Opin. Struct. Biol.* **30**, 63–70 (2015).
25. Shoji, K., Suzuki, Y., Sugano, S., Shimada, T. & Katsuma, S. Artificial “ping-pong” cascade of PIWI-interacting RNA in silkworm cells. *RNA* **23**, 86–97 (2017).
26. Matsumoto, N. *et al.* Crystal structure of silkworm PIWI-clade Argonaute Siwi bound to piRNA. *Cell* **167**, 484–497.e9 (2016).
27. Saxe, J. P., Chen, M., Zhao, H. & Lin, H. Tdrkh is essential for spermatogenesis and participates in primary piRNA biogenesis in the germline. *EMBO J.* **32**, 1869–1885 (2013).
28. Ding, D. *et al.* PNLD1 is essential for piRNA 3' end trimming and transposon silencing during spermatogenesis in mice. *Nat. Commun.* **8**, 819 (2017).
29. Handler, D. *et al.* A systematic analysis of *Drosophila* TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors. *EMBO J.* **30**, 3977–3993 (2011).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We are grateful to T. Mannen for preparing materials for mass spectrometry, T. Suzuki for comments on our *in vitro* Zuc processing assays and Y. Ono for support with the bioinformatics. We also thank S. Ohnishi for technical assistance and other members of the Siomi laboratories for discussions and comments on the manuscript. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan to K.M.N., Y.W.I., Y.M., H.S. and M.C.S. R.M. is supported by CREST, the Japan Science and Technology Agency. T.Ka. was supported by grants from the New Energy and Industrial Technology Development Organization, Japan, and Translational Systems Biology and Medicine Initiative from the Ministry of Education, Culture, Sports, Science and Technology of Japan. T.Ko. is a recipient of Molecular Dynamics for Antibody Drug Development, First Program Grant from the Japan Society of Promotion of Science.

**Author Contributions** K.M.N. generated monoclonal antibodies and performed biochemical analyses of piRNAs, int-piRNAs and piRNA factors. K.S. carried out *in vitro* experiments with help from R.M. H.Y. performed protein–protein interaction analyses. Y.M. performed immunofluorescence analyses. Y.W.I. performed bioinformatics analyses. T.Ka. and T.Ko. performed LC–MS/MS analysis. M.C.S. designed the experiments with other authors, supervised and discussed the work, and wrote the manuscript. H.S. discussed and supervised the study. All authors commented on the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.C.S. ([siomim@bs.s.u-tokyo.ac.jp](mailto:siomim@bs.s.u-tokyo.ac.jp)).

**Reviewer Information** *Nature* thanks J. Brennecke, S. Chameyron and the other anonymous reviewer(s) for their contribution to the peer review of this work.



## METHODS

**Production of monoclonal antibodies.** Monoclonal antibodies were produced essentially as described previously<sup>14</sup>. Mice were immunized with bacterially produced glutathione S-transferase (GST)-tagged Papi (amino acids 430–629), GST-Trim (amino acids 1–240) or GST-Zuc (amino acids 63–206). The Papi, Trim and Zuc cDNAs corresponding to the particular regions were obtained by RT-PCR using total RNAs from BmN4 cells. The PCR primers used are summarized in the 'Oligonucleotides' section.

**Plasmid construction.** pIB-myc and pIB-3xFlag were generated using a pIB vector (Thermo Fisher Scientific). Vectors to express Myc-Siwi and Papi-Flag were generated by inserting Siwi and Zuc cDNAs into pIB-myc and pIB-3xFlag, respectively. Vectors for expressing the Papi-Flag KH mutant, the Papi-Flag siRNA-resistant mutant, the Flag-Siwi RK mutant and the Flag-Ago3 RK mutant were generated by inverse PCR using expression vectors for Papi-Flag (this study), Flag-Siwi<sup>14</sup> and Flag-Ago3<sup>14</sup> as templates. The PCR primers used are summarized in the 'Oligonucleotides' section.

**RNAi and transgene expression.** Double-stranded RNAs (dsRNAs) were produced by *in vitro* T7 transcription, followed by annealing in water. The PCR primers to generate dsRNAs are summarized in the 'Oligonucleotides' section. BmN4 cells were transfected with 2 µg of dsRNAs (per 3 × 10<sup>5</sup> cells) using FuGENE HD (Promega). To express proteins exogenously in BmN4 cells, the cells were transfected with 2 µg of plasmids (per 6 × 10<sup>5</sup> cells) using FuGENE HD.

**Western blot analysis.** Western blotting was carried out as described previously<sup>14</sup>. The Y12 antibody was a gift from G. Dreyfuss. Anti-HSP60 (StressMarq Biosciences), anti-tubulin (Developmental Studies Hybridoma Bank), anti-DDDDK-tag (Flag) (MBL), and anti-Myc (Sigma) antibodies were purchased. **qRT-PCR.** qRT-PCR was performed as described previously<sup>30</sup>. The PCR primers used are summarized in the 'Oligonucleotides' section.

**Immunoprecipitation and RNA isolation.** Immunoprecipitation of Flag-Siwi and Flag-Ago3 was carried out essentially as described previously<sup>14</sup>. BmN4 whole and mitochondrial lysates (Fig. 2c, f and Extended Data Fig. 6d) were prepared in binding buffer (30 mM HEPES (pH 7.4), 150 mM potassium acetate, 5 mM magnesium acetate, 5 mM DTT, 0.5% Triton X-100, 2 µg ml<sup>-1</sup> pepstatin, 2 µg ml<sup>-1</sup> leupeptin and 0.5% aprotinin) and incubated with anti-Papi, anti-Siwi or anti-Flag antibody bound to Dynabeads (Invitrogen) at 4 °C for 2 h. The beads were washed four times with binding buffer. BmN4 whole (Fig. 1h and Extended Data Fig. 5b) and mitochondrial (Extended Data Fig. 5c) lysates were prepared in binding buffer containing 500 mM sodium chloride and incubated with anti-Papi or anti-Siwi antibody. The beads were washed twice with binding buffer containing 500 mM sodium chloride, and then twice with binding buffer. RNAs were eluted from the beads by phenol-chloroform after protease K treatment and precipitated with ethanol. RNA <sup>32</sup>P-labelling<sup>14</sup>, β-elimination<sup>22</sup> and northern blotting<sup>12</sup> were carried out as described previously. RNAs were crosslinked to Hybond-N<sup>+</sup> (GE Healthcare) by UV irradiation. The sequences of the RT3-1 probe and siRNA used in the β-elimination experiment are described in 'Oligonucleotides'.

**Generation of PIWI-associated small RNA libraries.** Immunoprecipitation of Flag-Siwi and Flag-Ago3 was carried out as described previously<sup>14</sup>. RNAs were eluted from the immunoprecipitates by phenol-chloroform after protease K treatment, and precipitated with ethanol. RNAs of 23–35 nucleotides in length were eluted from the gels and used to generate small RNA libraries<sup>14</sup>.

**Analysis of PIWI-associated small RNA sequences.** Libraries were sequenced using Illumina MiSeq (single-end, 51 cycles). For Siwi-associated small RNAs, a total of 4,547,701 reads were obtained from the control sample, 3,054,386 reads from the Trim-knockdown sample and 4,083,584 reads from the Nbr-knockdown sample. For Ago3-associated small RNAs, a total of 3,855,720 reads were obtained from the control sample, 3,702,496 reads from the Trim-knockdown sample and 3,292,910 reads from the Nbr-knockdown sample. The analysis of small RNAs was performed as described previously<sup>14</sup>. In brief, adaptor sequences were removed from the obtained reads, and the reads in the range of 23–35 nucleotides were used for further analysis (89–94% of the sequenced reads were in this size range). The reads were mapped to the silkworm reference genome (downloaded from the Silkworm Genome Research Program Database; <http://sgp.dna.affrc.go.jp/data/integretedseq.txt.gz>) using Bowtie<sup>31</sup>, allowing no mismatches. Genome mapped reads were extracted and aligned to 121 *B. mori* transposon consensus sequences (a gift from S. Kawaoka) using Bowtie<sup>31</sup>, allowing up to two mismatches. Using transposon-mapped reads, the length distribution was calculated. Sequence logos were calculated using the motifStack R package. The sequences were aligned to the 5' end upon the calculation of sequence logos. Strand bias and frequency (reads per million) of small RNA reads were calculated for 70 transposon consensus sequences with higher amount of mapped Siwi and Ago3 piRNAs (reads per million) in control sample, and heat maps were depicted using Java TreeView software<sup>32</sup>.

**Dephosphorylation treatment.** Immunopurified Papi was incubated with Antarctic Phosphatase (NEB) at 37 °C for 30 min for dephosphorylation.

**Immunofluorescence.** Immunofluorescence was carried out essentially as described previously<sup>14</sup>. Anti-Flag antibody and Alexa Fluor 488 goat anti-mouse IgG antibody (Invitrogen) were used as primary and secondary antibodies, respectively.

**Trypsin digestion and LC-MS/MS analysis.** The method for trypsin digestion of protein has been described previously<sup>33</sup>. Liquid chromatography–tandem mass spectrometry (LC-MS/MS) analysis was performed using an LTQ Orbitrap XL electron transfer dissociation (ETD) mass spectrometer (Thermo Fisher Scientific). The methods used for LC-MS/MS were slightly modified from those described previously<sup>34</sup>. The mass spectrometer was operated in a data-dependent acquisition mode in which the mass spectrometry acquisition with a mass range of *m/z* 420–1,600 was automatically switched to MS/MS acquisition under the automated control of Xcalibur software. The top four precursor ions in an MS scan were selected by Orbitrap, with resolution *R* = 60,000 and in subsequent MS/MS scans by ion trap in the automated gain control (AGC) mode in which the AGC values were 5.00 × 10<sup>5</sup> and 1.00 × 10<sup>4</sup> for full MS and MS/MS, respectively. To analyse dimethylation sites, ETD was used.

**Database searching and protein identification.** Database searches were performed using the MASCOT 2.5.1 search engine (Matrix Science) against the UniProtKB\_2016-8 database (selected for *B. mori*), assuming trypsin as the digestion enzyme and allowing for trypsin specificity of up to four missed cleavages. The database was searched with a fragment ion mass tolerance of 0.60 Da and a parent ion tolerance of 3.0 p.p.m. The iodoacetamide derivative of cysteine was specified as a fixed modification and methylation of arginine, oxidation of methionine, dimethylation of arginine and acetylation of N termini were specified as variable modifications. Scaffold (version Scaffold\_4.7.5; Proteome Software) was used to validate MS/MS-based peptide and protein identifications. We accepted the peptide identifications when the Peptide Prophet algorithm<sup>35</sup> specified probabilities at >95.0%. Sequence coverage was defined as the percentage of the protein in the identified peptide sequence.

**CLIP.** CLIP was performed as described previously<sup>36</sup>. Dephosphorylation and RNA radiolabelling with <sup>32</sup>P were performed using T4 polynucleotide kinase.

**Rescue assay.** BmN4 cells were transfected with 500 pmol siRNA duplex (per 1 × 10<sup>6</sup> cells) using Cell Line Nucleofector Kit L (Lonza) and incubated at 27 °C for 72 h. The sequences of siRNAs are presented in 'Oligonucleotides'. After RNAi, cells were transfected with 2 µg of Papi-Flag plasmid using FuGENE HD and incubated at 27 °C for 72 h. Cells were then transfected with 2 µg Myc-Siwi plasmid and incubated at 27 °C for 24 h. BmN4 lysates were prepared in binding buffer (30 mM HEPES (pH 7.4), 150 mM potassium acetate, 5 mM magnesium acetate, 5 mM dithiothreitol (DTT), 0.1% Tergitol-type NP-40, 2 µg ml<sup>-1</sup> pepstatin, 2 µg ml<sup>-1</sup> leupeptin and 0.5% aprotinin) containing 500 mM sodium chloride, and incubated with anti-Myc antibody bound to Dynabeads. The beads were washed twice with binding buffer containing 500 mM sodium chloride and then twice with binding buffer. RNAs were eluted from the beads by phenol-chloroform after protease K treatment and precipitated with ethanol. RNA radiolabelling was carried out as described previously<sup>14</sup>.

**Sucrose density gradient centrifugation.** Mitochondria were prepared as described previously<sup>37</sup>. The mitochondrial pellet was resuspended in gradient buffer (30 mM HEPES (pH 7.4), 100 mM potassium acetate, 1 mM DTT, 4 mM magnesium acetate, and 1% Tergitol-type NP-40), and centrifuged at 14,000g at 4 °C for 30 min. The supernatant was placed at the top of a 10–40% sucrose gradient and centrifuged in a Beckman MLS-50 rotor at 178,000g at 4 °C for 16.5 h.

**Preparation of Papi-associated int-piRNA libraries.** RNAs were eluted from the Papi immunoprecipitates by phenol-chloroform after protease K treatment and precipitated with ethanol. RNAs were resolved by denaturing PAGE, and 65–100-nucleotide-long RNAs were eluted from gels. The libraries were generated as described previously<sup>14</sup>.

**Sequence analysis of Papi-associated int-piRNAs.** Libraries were sequenced using Illumina MiSeq (single-end, 111 cycles). A total of 10,584,873 reads were obtained from the control sample and 7,932,439 reads were obtained from the Zuc-knockdown sample. Adaptor sequences were removed from the obtained reads, and the reads were mapped to the silkworm reference genome (downloaded from the Silkworm Genome Research Program Database; <http://sgp.dna.affrc.go.jp/data/integretedseq.txt.gz>) using Bowtie<sup>31</sup>, allowing no mismatches. Genome mapped reads were extracted and aligned to *B. mori* transposon consensus sequences using Bowtie<sup>31</sup>, allowing up to two mismatches. To reduce the effect of contaminant reads, the reads mapped to transposon consensus sequences were used for further analysis. Sequence logos were calculated using the motifStack R package (<http://www.bioconductor.org/packages/release/bioc/html/motifStack.html>). The sequences were aligned to the 5' end upon the calculation of sequence logos. Calculations of the distance between intermediates and piRNA-phasing analysis were performed as previously described<sup>6</sup>, and complementarity of piRNAs and adjacent intermediate pairs was calculated using an in-house script.

The phasing analysis and piRNA-intermediate complementarity calculation were performed using each pair of intermediates preserving the read count. Siwi and Ago3 piRNA sequences were obtained from previously published data<sup>14</sup> (GEO accession GSE58221).

**Recombinant protein preparation.** The cDNA encoding Zuc excluding MLS (residues 29–206) was cloned into pCold-GST vector. The Zuc(H141A) mutant was generated from pCold-Zuc by inverse PCR. The protein was purified using glutathione-Sepharose (GE Healthcare) with purification buffer (20 mM Tris-HCl (pH 8.0), 150 mM sodium chloride and 1 mM DTT). The proteins were treated with HRV3C protease (GE Healthcare) to remove the GST tag and further purified by Enrich S (Bio-Rad). To yield pIZ-3xFlag-Siwi, 3xFlag-Siwi was amplified from pIB-3xFlag-Siwi<sup>14</sup> by PCR and inserted into pIZ vector (Thermo Fisher). Flag-Siwi was immunoprecipitated by ANTI-FLAG M2 Affinity Gel (Sigma) from BmN4 cells and eluted by 500 ng  $\mu\text{l}^{-1}$   $\times$  Flag peptide (Sigma).

**In vitro processing assay.** The synthesized RNAs (GeneDesign) were <sup>32</sup>P-labelled at their 5' end. The sequences of RNAs are indicated in 'Oligonucleotides'. Radiolabelled int-piRNAs (10<sup>4</sup> c.p.m.) were incubated with 0.1  $\mu\text{g}$  of recombinant Zuc in 20  $\mu\text{l}$  of buffer A<sup>18</sup> at 26 °C for 30 min. A total of 1  $\mu\text{l}$  of 1  $\mu\text{M}$  radiolabelled int-piRNAs and 200 ng of purified Flag-Siwi were mixed at 26 °C for 30 min in 20  $\mu\text{l}$  of the loading buffer (30 mM HEPES (pH 7.4), 100 mM potassium acetate, 2 mM magnesium acetate, 20 mM creatine monophosphate, 1 mM ATP, 0.15  $\mu\text{M}$  <sup>1</sup> creatine phosphokinase, 1 mM DTT and 0.5  $\mu\text{M}$  <sup>1</sup> RNasin (Promega)). After loading, the Flag-Siwi-containing mixture was incubated with Dynabeads (with anti-Siwi antibody or Papi immunoprecipitation product) at 4 °C for 1 h. The beads were then washed five times with binding buffer. RNAs were isolated from the beads with phenol-chloroform and precipitated with ethanol. They were then resolved by denaturing PAGE. Immunopurified Papi complexes were incubated with 1  $\mu\text{g}$  of recombinant Zuc in 30  $\mu\text{l}$  of buffer A<sup>18</sup> containing 2.5 mM EGTA and 0.1  $\mu\text{M}$  <sup>1</sup> RNasin at 27 °C for 0–180 min. RNAs were isolated from the beads with phenol-chloroform and precipitated with ethanol, after which they were transferred to Hybond-N (GE Healthcare). Next, RNAs were chemically crosslinked to membrane using 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide as described previously<sup>14</sup>. The sequence of probe RT3-1 is indicated in the 'Oligonucleotides' section.

**Oligonucleotides.** Primers used for producing constructs for Myc-Siwi, Papi-Flag, siRNA-resistant Papi, Papi KH mutant, Flag-Siwi RK mutant, Flag-Ago3 RK mutant and pIZ-3xFlag-Siwi are listed below.

Myc-L (left, forward): 5'-TTCGAATTTAAAGCTCACCATGGGA GAGCAGAACTGATC-3', Myc-R (right, reverse): 5'-CTGCAGGAATTCGAT CCGGGTACCAAGCTTGCTAG-3'; 3xFlag-L: 5'-GGCCCGCGGTTTCAAGA CTACAAAGACCAT-3', 3xFlag-R: 5'-AGTCAGATAAACTCAGATATCCT TGTCTATC-3'; Myc-Siwi-L: 5'-TGCAGCCCAGCGCTGGATCCATG TCAGAACCGAGAGGTAG-3', Myc-Siwi-R: 5'-CGAACC CGGGCCCTTTAG AGGAAATATAAAGTTT-3'; Papi-Flag-L: 5'-TCGAATTTAAAGCTTATGT CATTGAACACAAAATT-3', Papi-Flag-R: 5'-GAACCGCGGGCCCTCCTTT TCAAAAGCGGACTTAC-3'; Papi siRNA-res-L: 5'-CACCAAGTCAACAGA TAAAGTTGTGAGCA-3', Papi siRNA-res-R: 5'-ACTCCATTTGACGCCG GCGCCATCGGAT-3'; Papi KH I69N mut-L: 5'-TCCACCAACAAAGGACC TTCAGAAGAAATCT-3', Papi KH I69N mut-R: 5'-GCCATTGCGACCAAT CAGAGCTGGAACAAT-3'; Papi KH V142N mut-L: 5'-GAGAACACAATGA TATTAGCCATCGCAGT-3', Papi KH V142N mut-R: 5'-ACCTCCAGATCCAATT ATTCTCCCAACAAGA-3'; Siwi RK mut1-L: 5'-CAGAACCAGAGGTAAAG GAAAAGCTAAAGGAAAGCTGTGTAAGGGTGGTATGGAGGC-3', Siwi RK mut1-R: 5'-GCCTCCATCACCACCTTACCAGCCTTTCTCTTATAGCTTT TCCTTACCTCTCGGTTCTG-3'; Siwi RK mut2-L: 5'-CGTAGTTGGCAAG GGCTCTAAAAAAGGGGTGGAAAAGTCTTCTCTG-3', Siwi RK mut2-R: 5'-CAGGAAGGACTTTTCCACCCCTTTTATAGAGCCCTTGCCAACTACG-3'; Ago3 RK mut1-L: 5'-CCAGGCAAAAGGCAAGGGGAAAGCTTAGCC-3', Ago3 RK mut1-R: 5'-GGGCTAAGCTTTTCCCTTGCCCTTTGCTTG-3'; Ago3 RK mut2-L: 5'-GTATAGGCGGTAAAGGAAAGGCAGCAGCATTG-3', Ago3 RK mut2-R: 5'-CAATGCTGCTGCCTTTTCCCTTACCGCCTATAC-3'; Ago3 RK mut3-L: 5'-CAGCTGGAATCGGAAAGGATTCAAATTCG-3'; Ago3 RK mut3-R: 5'-GCAATTTGAATCCCTTTCCAGTTCCAGCTG-3'; 3xFlag-Siwi-L: 5'-TCGAATTTAAAGCTTACCATGGACTACAAAGACCATGACGGTG-3', 3xFlag-Siwi-R: 5'-GAACGAGAAACGTAAGTTTATAGAGGAAATATA AAGTTTCATTC-3'.

Primers used for producing constructs for GST-Papi, GST-Trim and GST-Zuc are: GST-Papi-L: 5'-TGGGATCCCCGAATTCAGGACAAAGAGATACCTGG-3', GST-Papi-R: 5'-GGCCGCTCGAGTCGACTCACTTTTCAAAAGCGGACT-3'; GST-Trim-L: 5'-TGGGATCCCCGAATTCATGGATATACCAAGAAAA-3', GST-Trim-R: 5'-GGCCGCTCGAGTCGACTTAGTTATCTTCCAGTA

TTGCTA-3'; GST-Zuc-L: 5'-TACCCTCGAGGGATCCACAAAAAGCATG GAC-3', GST-Zuc-R: 5'-CGACAAGCTTGAATTTTAACTGGTTATTGG-3'.

Primers used for producing constructs for pCold-Zuc and the Zuc(H141A) mutant are: pCold-Zuc-L: 5'-TACCCTCGAGGGATCCAGAAGAAGAA AGAA-3', pCold-Zuc-R: 5'-CGACAAGCTTGAATTTTAACTGGTTATTGG-3'; Zuc(H141A)-L: 5'-GCCCCACAAGTTCTGCATAATAGATG-3', Zuc(H141A)-R: 5'-CATTAGATTTGTAGACTTCATCCAG-3'.

Primers used to generate templates for dsRNA production are: T7-dsLuc-L: 5'-TAATACGACTCACTATAGGGGGAGAGCAACTGCATAAGGC-3', T7-dsLuc-R: 5'-TAATACGACTCACTATAGGGTCCCTATCGAA GGACTCTGG-3'; T7-dsPapi-L: 5'-TAATACGACTCACTATAGGGTCT TAGTGACATTCCTGGTA-3', T7-dsPapi-R: 5'-TAATACGACTCACTATAGG GAGCATCCCTGGCTTGGAAAC-3'; T7-dsTrim-L: 5'-TAATACGACTCAC TATAGGGTTTCAAGTTTCAAAATGGT-3', T7-dsTrim-R: 5'-TAATACGAC TCACTATAGGGAGCGAAGAATTCATACAAAT-3'; T7-dsZuc-L: 5'-TAATAC GACTCACTATAGGGATGGCAGTAACTCTTAGTAA-3', T7-dsZuc-R: 5'-TAATACGACTCACTATAGGGTCCAGTTCAATGACCTGC-3'; T7-dsNbr-L: 5'-TAATACGACTCACTATAGGGGACAAATGTTATGGCATTGG-3', T7-dsNbr-R: 5'-TAATACGACTCACTATAGGGCAATCCATTTCAAGGCTTCA-3'; T7-dsSiwi-L: 5'-TAATACGACTCACTATAGGGGCAAAATAATGTCAAATACCC-3', T7-dsSiwi-R: 5'-TAATACGACTCACTATAGGGACCACCATCAGTGACGGCAG-3'.

Sequences of siRNA duplexes used for RNAi (DNA: normal font, RNA: underlined) are: Luc siRNA duplex: 5'-CGUACGCGGAAUUCUUCGATT-3' and 5'-UCGAAGUAUUCGCGUACGTT-3'; Papi siRNA duplex: 5'-GGUCG AAAGUCCUAAAAGUTT-3' and 5'-ACUUUUAGGACUUUCGACCTT-3'.

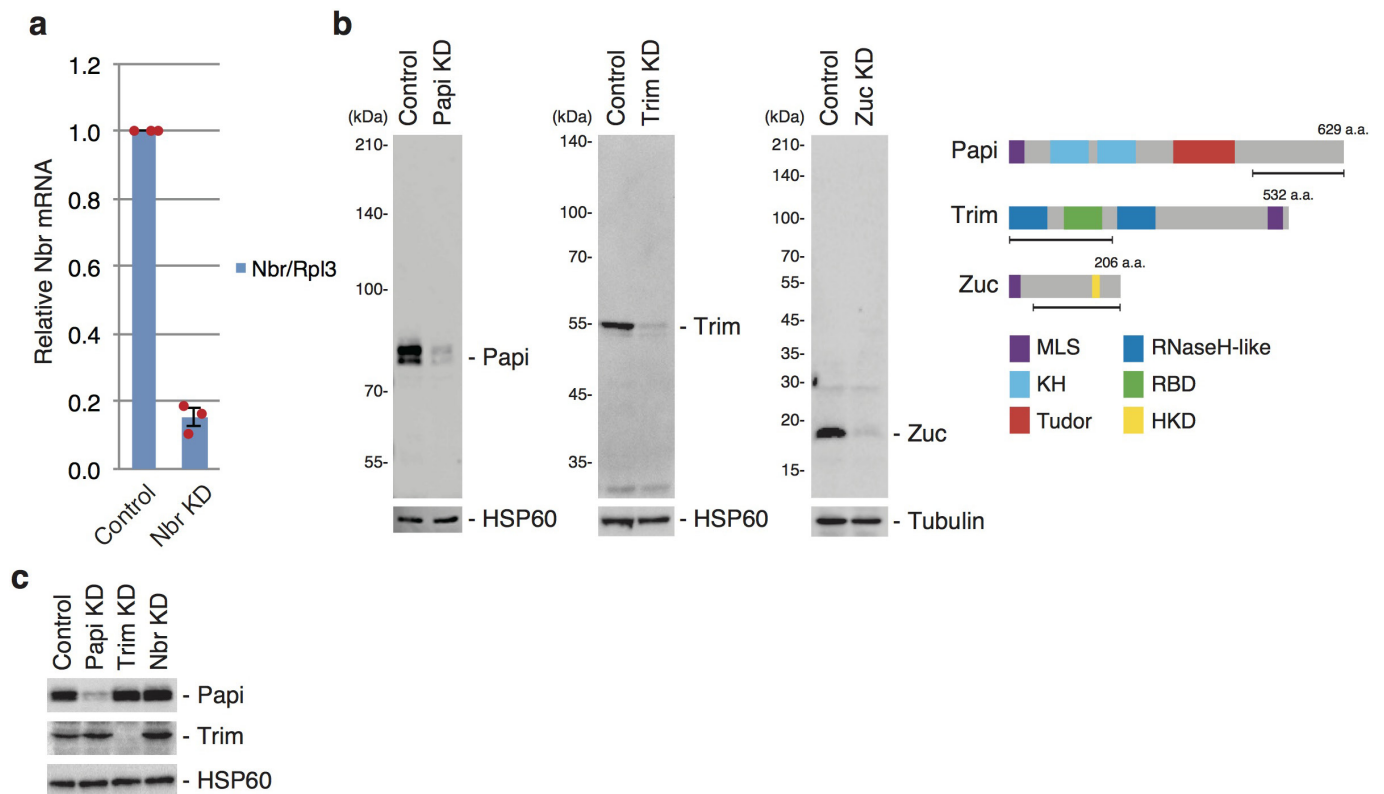
Sequence of RT3-1 probe used for northern blot is: 5'-ACCAGCCGATCGTC ATCGCATCCGTTTA-3'. Sequence of siRNA used for  $\beta$ -elimination is: 5'-UGGUCU GCCUAAAGGUGUCGUCUGC-3'. Primers used for qPCR are: Nbr-L: 5'-ACAGCCAGTTCAAAATAGTTATTGC-3', Nbr-R: 5'-TTGACCACAGT ATTACACAGAACT-3'; Rpl3-L: 5'-GGTGTACCAAGGGCAAAG-3', Rpl3-R: 5'-AGGATGCCAAGCTCCAATGC-3'.

Sequences of RNA used for processing assays are: 1U29: 5'-UCAA A AACUAACGGAUUGGUUUCGAACAG-3'; 1U50: 5'-UCAA A AACUAACGGAU UGGUUCGAACAGUACCCGCCGACAGGUCCC-3'; 1U80: 5'-UCAA A AACUAACGGAUUGGUUUCGAACAGUACCCGCCGACAGGUCCC CUACCUGUCCCUAAUACUUGGACGCCGGG-3'.

**Statistics and reproducibility.** Experiments in Figs 1a, e, 2b–e and 4d were performed three times independently with similar results. Experiments in Figs 1b–d, f–h, 2a, f and 4a–c were performed twice independently with similar results. Experiments in Extended Data Figs 1b, c, 4b, 5b, c, f, g, 6a, b were performed three times independently with similar results. Experiments in Extended Data Figs 3a, 4a, c, d, 5a, d, e, h, 6d and 7a–h were performed twice independently with similar results. Experiments in Extended Data Fig. 3c were performed once. No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

**Data availability.** Gel source images for Figs 1a–h, 2a–f and 4a–d and Extended Data Figs 1b, c, 3a, 4a, b, d, 5a–h, 6a, b, d and 7a–h are available in Supplementary Fig. 1. All other data supporting the findings of this study are available from the corresponding author upon reasonable request. All sequencing data that support the findings of this study were deposited in the NCBI Gene Expression Omnibus (GEO) with the GEO series accession number GSE107371.

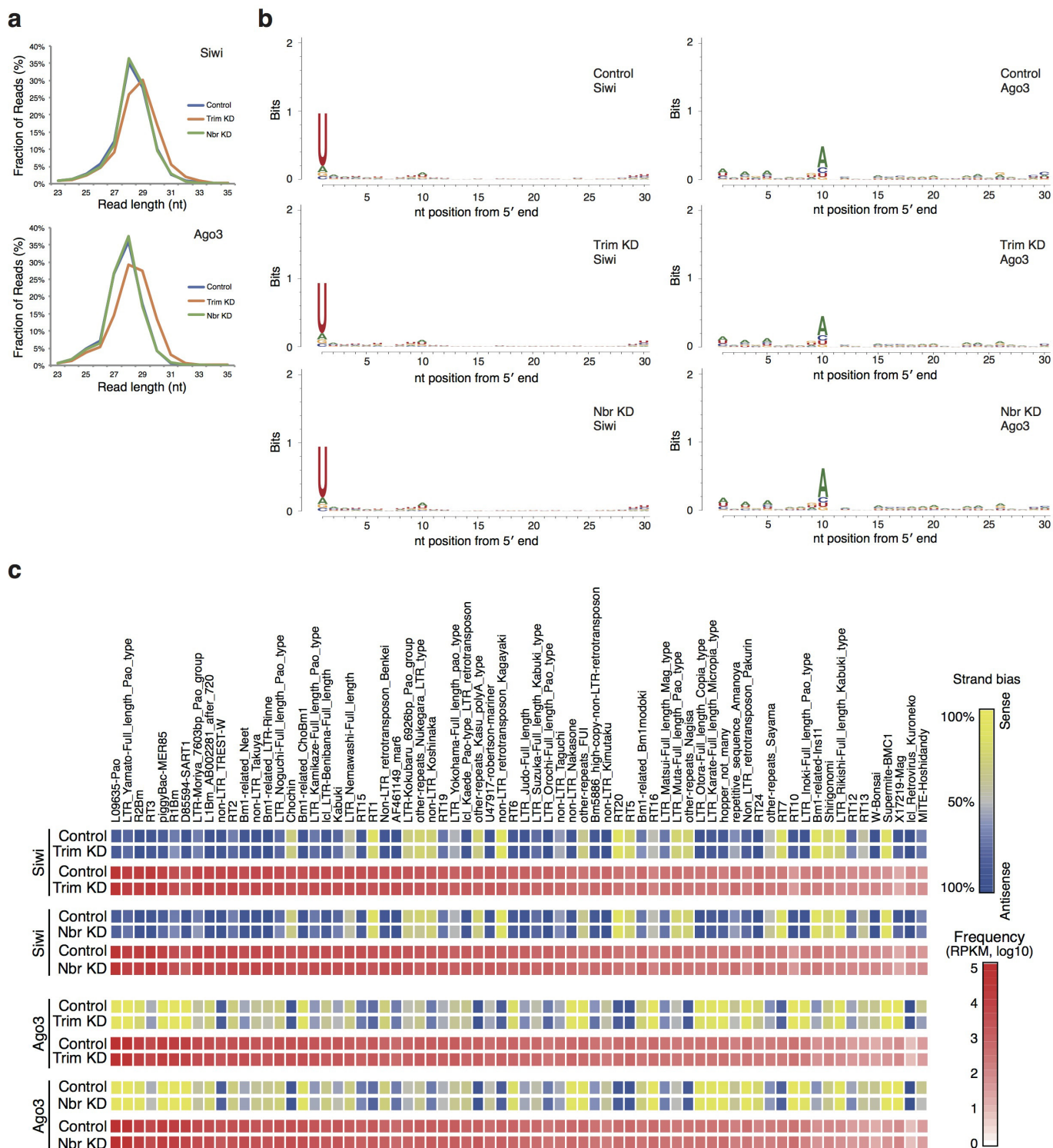
- Sumiyoshi, T. *et al.* Loss of *l(3)mbt* leads to acquisition of the ping-pong cycle in *Drosophila* ovarian somatic cells. *Genes Dev.* **30**, 1617–1622 (2016).
- Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
- Saldanha, A. J. Java Treeview extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
- Fujinoki, M. *et al.* Identification of 36-kDa flagellar phosphoproteins associated with hamster sperm motility. *J. Biochem.* **133**, 361–369 (2003).
- Fujii, K. *et al.* Fully automated online multi-dimensional protein profiling system for complex mixtures. *J. Chromatogr. A* **1057**, 107–113 (2004).
- Keller, B. O., Wang, Z. & Li, L. Low-mass proteome analysis based on liquid chromatography fractionation, nanoliter protein concentration/digestion, and microspot matrix-assisted laser desorption/ionization mass spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **782**, 317–329 (2002).
- Murota, Y. *et al.* Yb integrates piRNA intermediates and processing factors into perinuclear bodies to enhance piRISC assembly. *Cell Rep.* **8**, 103–113 (2014).
- Wieckowski, M. R., Giorgi, C., Lebiedzinska, M., Duszynski, J., & Pinton, P. Isolation of mitochondria-associated membranes and mitochondria from animal tissues and cells. *Nat. Protocols* **4**, 1582–1590 (2009).



**Extended Data Figure 1 | Production of monoclonal antibodies against Papi, Trim and Zuc, and analysis of depletion upon RNA interference treatment.** **a**, Quantitative PCR with reverse transcription (qRT-PCR) shows that *Nbr* was efficiently depleted by RNA interference (RNAi) in BmN4 cells. Data are mean  $\pm$  s.e.m. of three independent experiments. **b**, Western blotting shows the specificity of anti-*Papi*, anti-*Trim* and

anti-*Zuc* monoclonal antibodies raised in this study. HSP60 and tubulin were used as loading controls. The images show the domain structures of *Papi*, *Trim* and *Zuc*. Underlines indicate the antigen regions used for producing the monoclonal antibodies. **c**, Western blotting shows that *Papi* and *Trim* were efficiently depleted by RNAi in BmN4 cells.

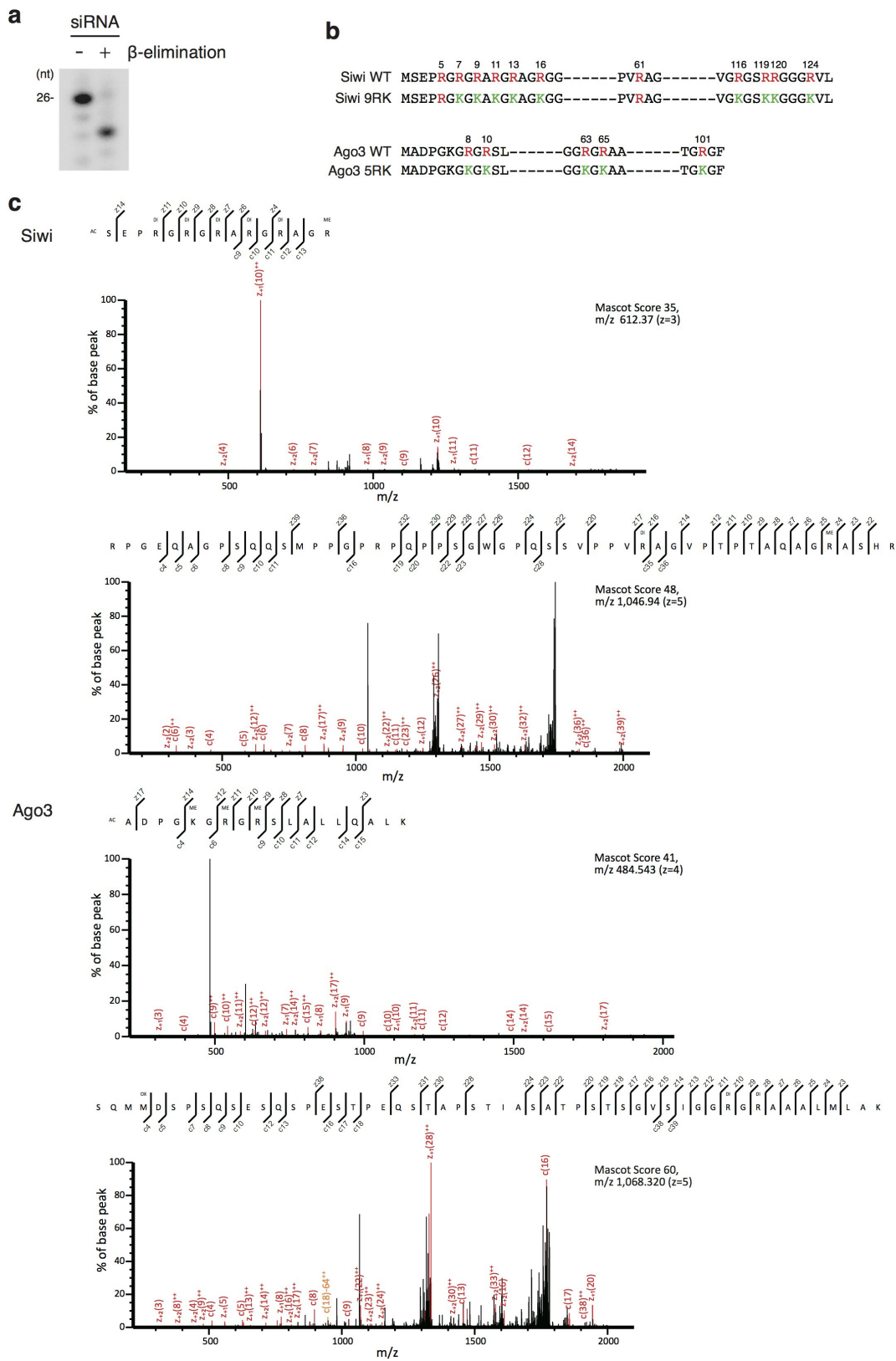




**Extended Data Figure 2 | Siwi/Ago3-associated small RNAs upon Trim or Nbr depletion.** **a**, Length distribution of transposon-mapped Flag-Siwi- and Flag-Ago3-associated piRNAs. piRNAs appear to be slightly longer when Trim was depleted. **b**, Sequence logos showing unaffected levels of

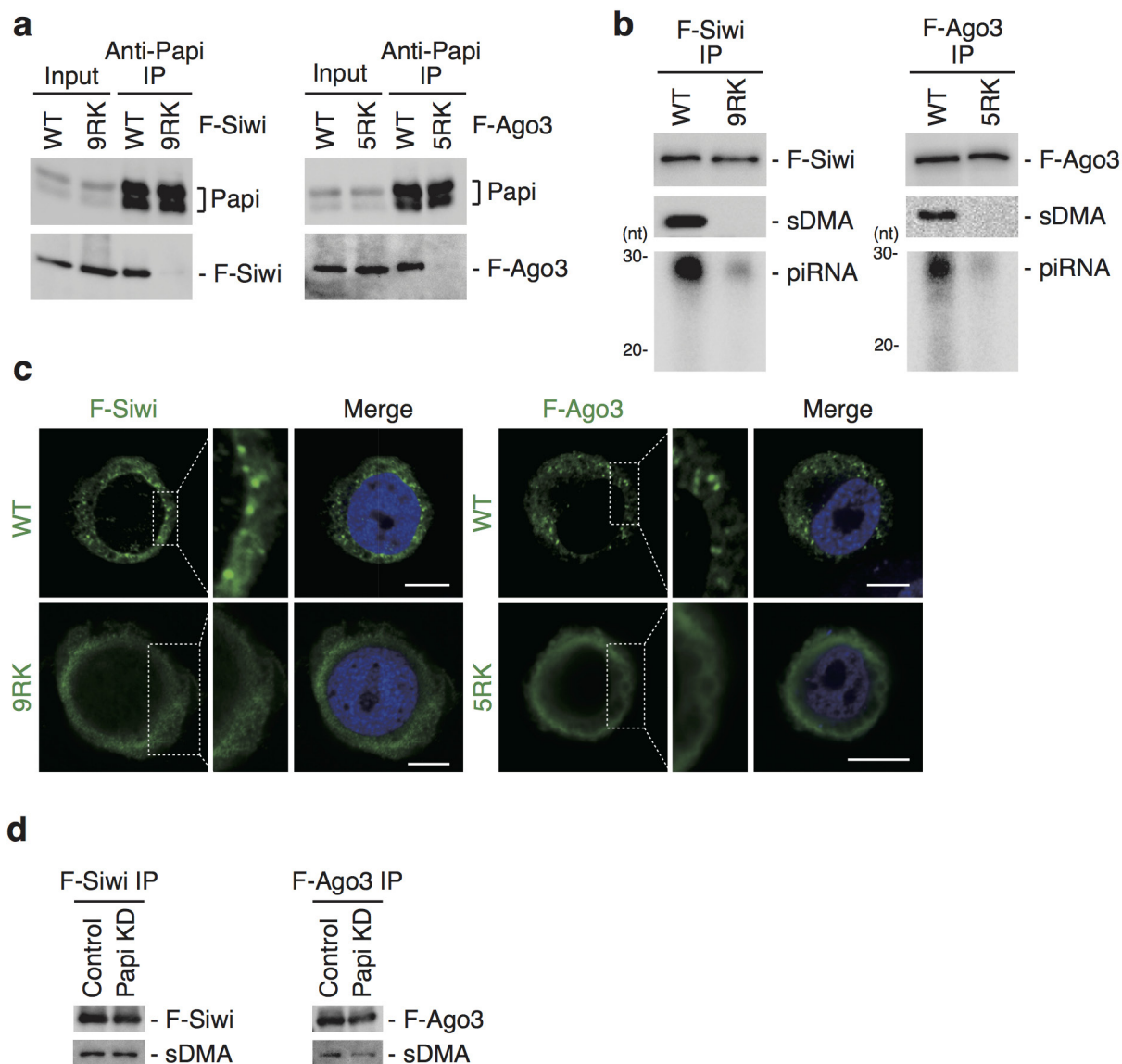
1U and 10A under Trim- or Nbr-depleted conditions. **c**, Strand bias and frequency of piRNAs mapped to each transposon consensus sequence. Depletion of Trim or Nbr has little effect on strand bias or the frequency of piRNAs mapped to each transposon consensus sequence.





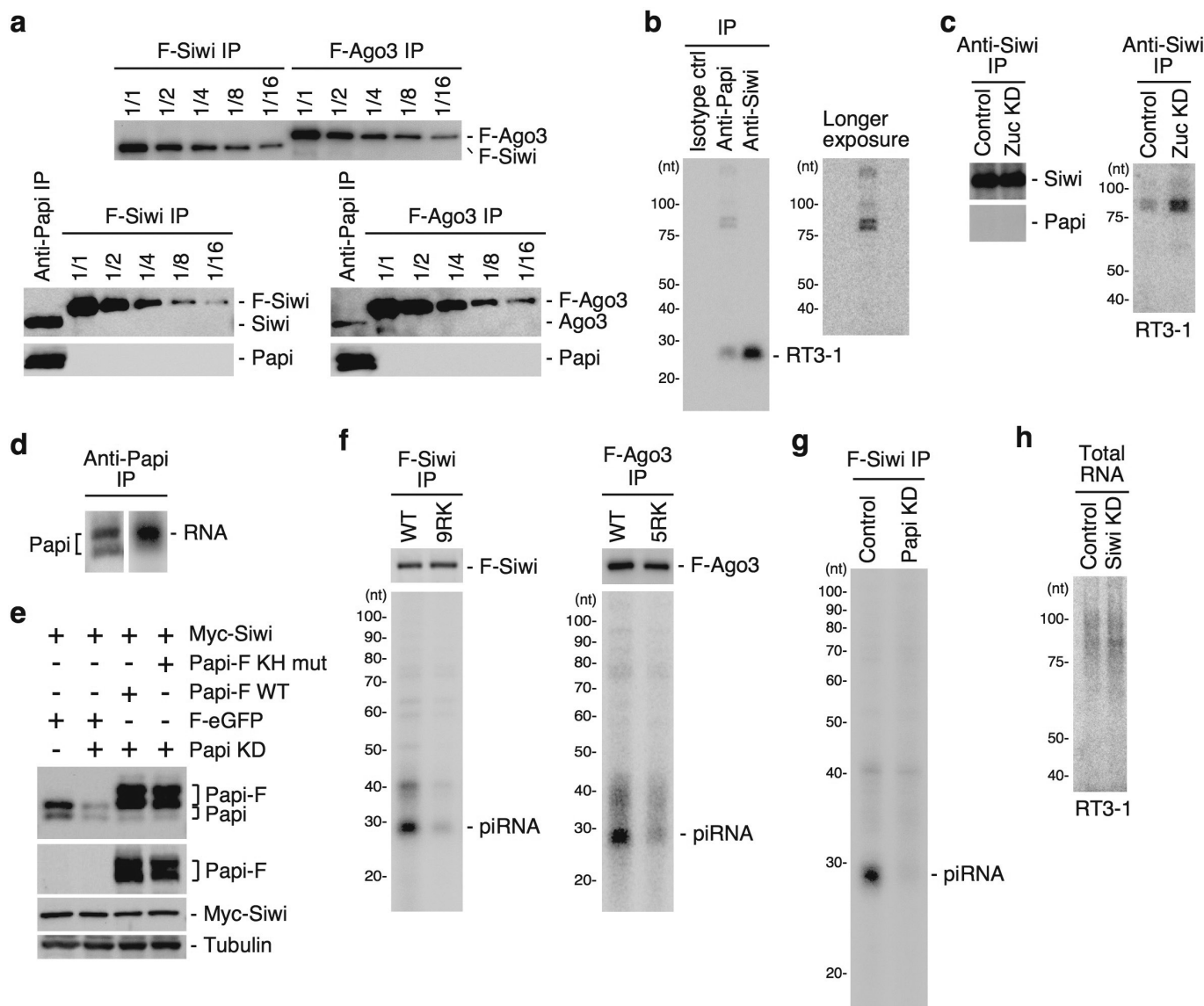
**a**, A synthesized short interfering RNA (siRNA) (26 nucleotides) was downshifted by  $\beta$ -elimination, indicating that this siRNA is not 2'-O-methylated. **b**, The amino acid sequences of the N-terminal regions of wild-type Siwi, the Siwi-9RK mutant, wild-type Ago3 and the Ago3-5RK mutant are shown. Arginine residues shown in red were determined to be

sDMAs in BmnN4 cells. Arginine residues mutated to lysines are shown in green. c, Representative ETD tandem mass spectra for Siwi and Ago3 peptides, which include arginine modifications. Ac, acetylation; Di, demethylation; Me, monomethylation. Charge,  $m/z$  and Mascot score are shown on the top right of each spectrum. All identified Siwi and Ago3 peptides are listed in Supplementary Table 1.



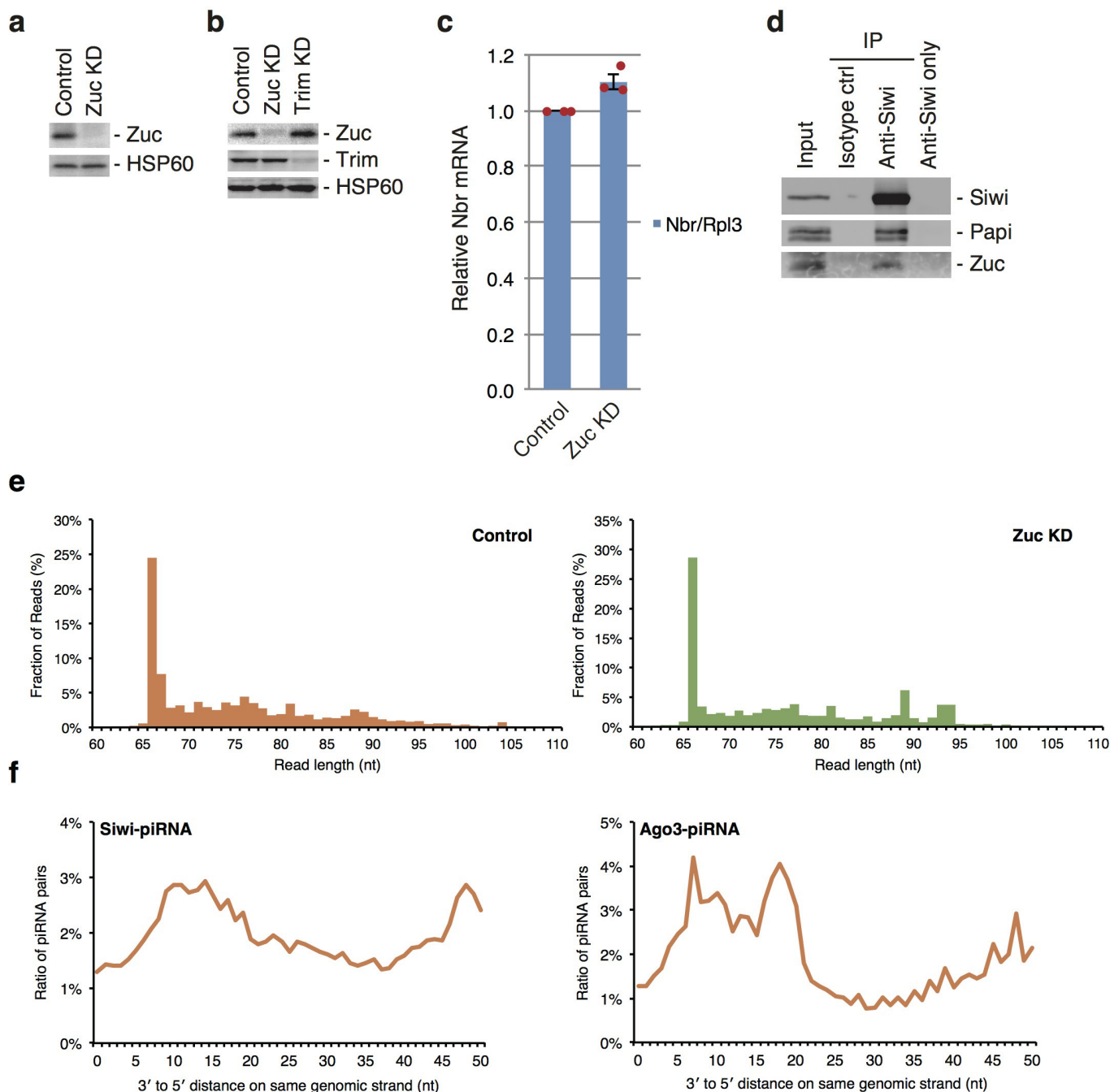
**Extended Data Figure 4 | Analysis of Siwi and Ago3 mutants.** **a**, Wild-type Flag-Siwi and Flag-Ago3, but not Flag-Siwi-9RK and Flag-Ago3-5RK mutants, are co-immunoprecipitated with Papi from BmN4 cells. **b**, Wild-type Flag-Siwi and Flag-Ago3, but not Flag-Siwi-9RK and Flag-Ago3-5RK mutants, are loaded with piRNAs in BmN4 cells. The middle (sDMA) shows that neither the Flag-Siwi-9RK nor Flag-Ago3-5RK

mutant reacts with the Y12 antibody, which specifically recognizes sDMA. **c**, Wild-type Flag-Siwi and Flag-Ago3, but not Flag-Siwi-9RK and Flag-Ago3-5RK mutants, are localized to nuage in BmN4 cells (shown in green). Blue (DAPI staining) indicates the location of the nucleus. Scale bars, 10  $\mu$ m. **d**, Papi depletion has little effect on sDMA modification of Flag-Siwi and Flag-Ago3 expressed in BmN4 cells.



**Extended Data Figure 5 | Papi complex analysis.** **a**, Top, Flag-Siwi and Flag-Ago3 expressed in BmN4 cells were immunoprecipitated with anti-Flag antibody and probed with anti-Flag antibody after sequential dilution. Bottom, Flag-Siwi and Flag-Ago3 immunoprecipitated from BmN4 cells (the same samples as in the top panel) were probed with anti-Siwi and anti-Ago3 antibodies, respectively. Siwi and Ago3 co-immunoprecipitated with Papi were simultaneously probed with anti-Siwi and anti-Ago3 antibodies, respectively. The Papi complex was equally divided into two fractions and each fraction was used for each blot. Examination of the signal intensity revealed that the amount of Siwi within the Papi complex was approximately equal to 1/1.6 volume of Flag-Siwi and that the amount of Ago3 within the Papi complex was approximately equal to 1/16 volume of Flag-Ago3. Comparison of the signal intensity on the top and bottom blots suggests that the ratio of abundance of Siwi and Ago3 in the Papi complex is 10:1. **b**, Northern blotting shows that the Papi complex contains RT3-1

int-piRNAs. **c**, Northern blotting shows that Siwi in a form associated with Papi on mitochondria binds RT3-1 int-piRNAs independently of Papi. The Siwi-int-piRNA association is maintained even after Zuc depletion. **d**, CLIP analysis shows that only the long form, but not the short form, of endogenous Papi in BmN4 cells interacts with RNA *in vivo*. **e**, Western blotting using anti-Papi (top) and anti-Flag (second from the top) antibodies shows that wild-type Papi-Flag and the KH mutant are equally expressed in BmN4 cells, in which endogenous Papi has been depleted by RNAi. Western blotting using anti-Myc (third from the top) shows that the levels of Myc-Siwi are approximately equal in the cells. Tubulin was used as a loading control (bottom). Both wild-type Papi-Flag and the KH mutant were mutated to be RNAi resistant. **f**, Flag-Siwi-9RK and Flag-Ago3-5RK mutants bind with little int-piRNA. **g**, Flag-Siwi binds with little int-piRNA in Papi-lacking BmN4 cells. **h**, Northern blotting shows that int-piRNAs are still present in Siwi-depleted BmN4 cells.

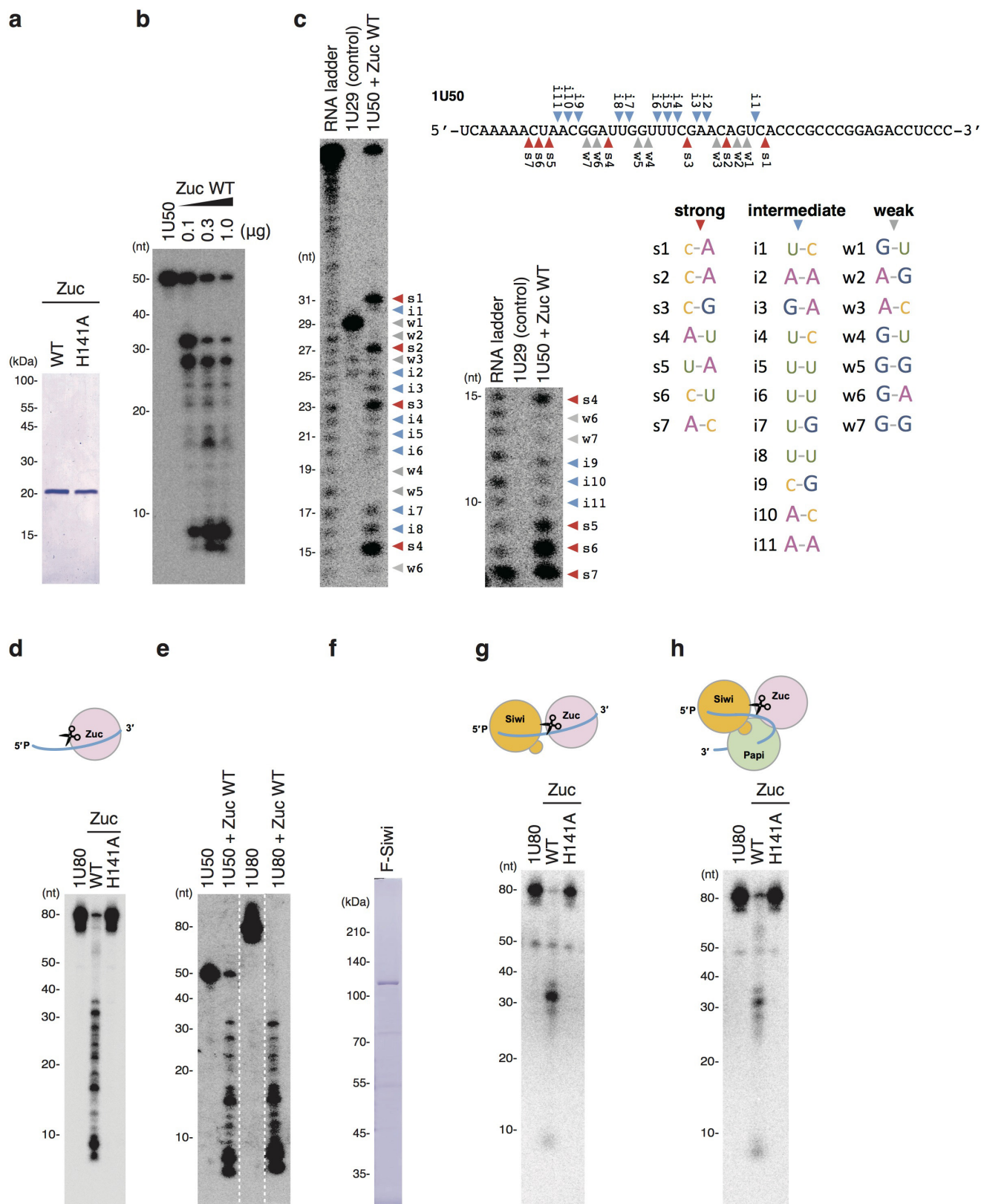


#### Extended Data Figure 6 | RNAi efficiency, Siwi-Papi-Zuc interaction and analysis of Papi-associated intermediates and piRNA phasing.

**a**, Western blotting shows that Zuc is efficiently depleted by RNAi. HSP60 is used as a loading control. **b**, Western blotting shows that Zuc and Trim are efficiently depleted by RNAi. Zuc and Trim depletion had little effect on the protein levels of Trim and Zuc, respectively. HSP60 is shown as a loading control. **c**, qRT-PCR shows that the level of *Nbr* is not affected by

Zuc depletion in BmN4 cells. Data are mean  $\pm$  s.e.m. of three independent experiments. **d**, Zuc and Papi are detected in the Siwi complex immunoprecipitated from the mitochondrial fraction of BmN4 cells. **e**, The size distribution of Papi-associated intermediates mapped to transposons. **f**, Analyses of phased piRNAs in Papi-associated intermediates. The distance between the 3' end of the upstream piRNA and the 5' end of the downstream piRNA on the same genomic strand is analysed.

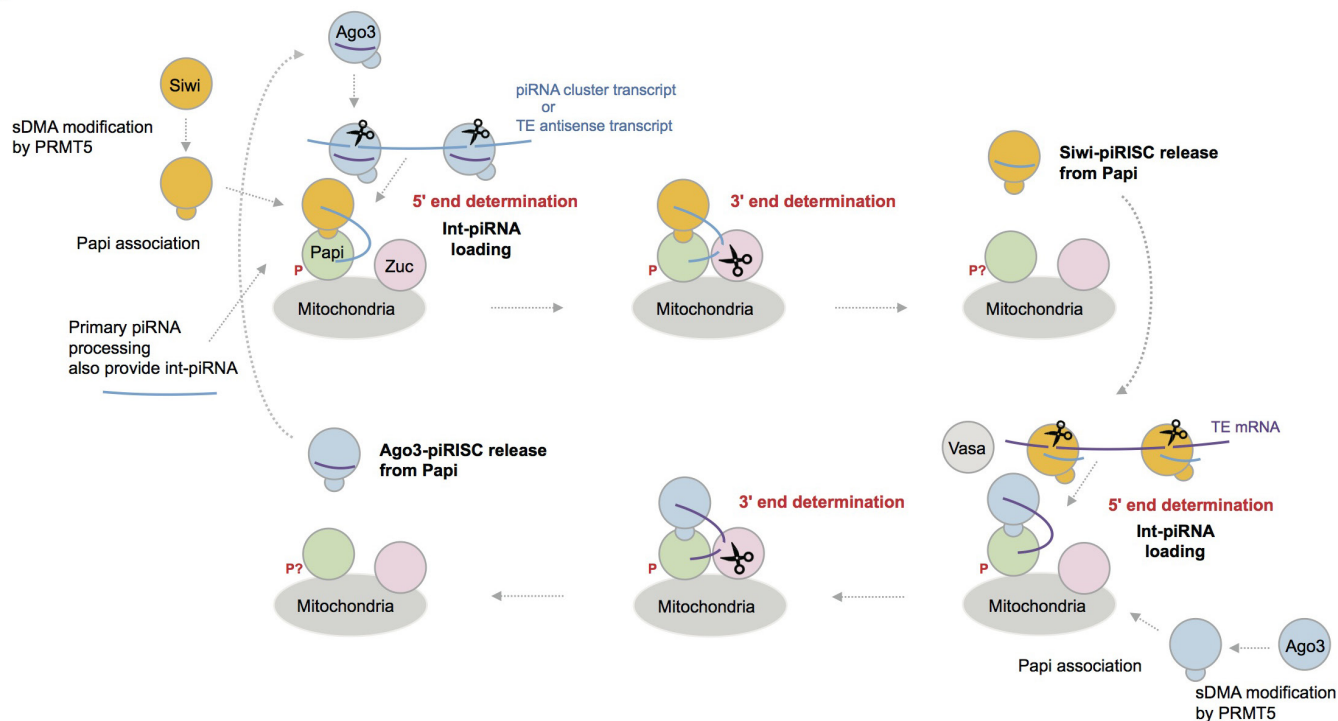




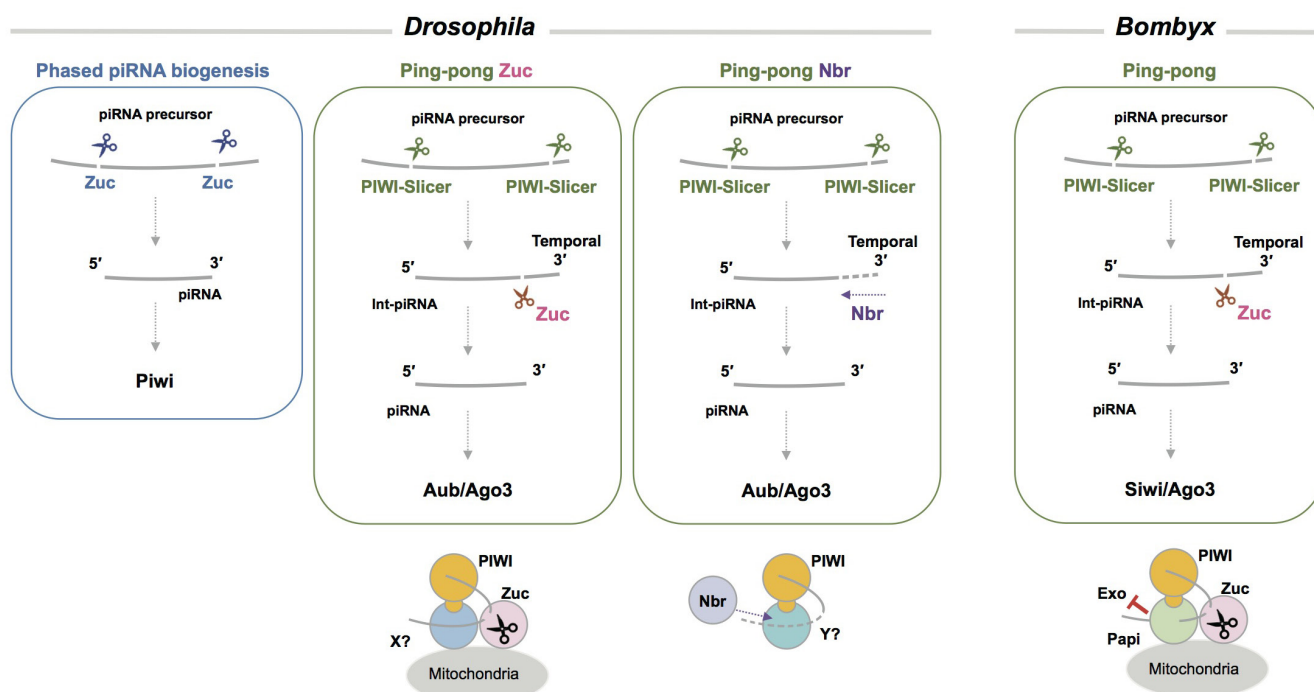
**Extended Data Figure 7 | Analyses of Zuc RNA cleavage.** **a**, Coomassie brilliant blue (CBB)-stained gel showing purified wild-type Zuc and the Zuc(H141A) mutant. **b**, Wild-type Zuc in **a** cleaves 1U50 in a dose-dependent manner. **c**, Detailed analyses of Zuc RNA cleavage. The left gel shows RNA ladders ranging from 14 to 50 nucleotides. The right gel shows RNA ladders ranging from 7 to 15 nucleotides. Relatively 'strong' RNA bands are indicated by red arrowheads (s1–s7). 'Intermediate' RNA bands are indicated by blue arrowheads (i1–i11). Relatively 'weak' RNA bands are indicated by grey arrowheads (w1–w7). 1U29 is an authentic RNA, the

sequence of which is identical to that of 1U50 RNA over 1–29 nucleotides from the 5' end. Classification of strong, intermediate and weak RNA bands is carried out in accordance with the intensity of each band. **d**, An 80-nucleotide RNA, 1U80, is cleaved by wild-type Zuc. **e**, Cleavage patterns of 1U50 and 1U80 by wild-type Zuc are compared. **f**, CBB-stained gel showing purified Flag-Siwi. **g**, 1U80 pre-loaded onto Flag-Siwi in **f** is cleaved by wild-type Zuc. **h**, The Siwi–1U80 RNA complex was first incubated with Papi, which was immunopurified using an anti-Papi antibody, and then treated with wild-type Zuc.

a



b



**Extended Data Figure 8 | A new model for piRNA biogenesis in *Bombyx*.** **a**, A model for the ping-pong cycle in *Bombyx*. Papi is localized on the surface of mitochondria through MLS, whereas KH domains are required for Papi to exhibit RNA-binding activity. The Papi Tudor domain and sDMA modification of Siwi and Ago3 are required for the Siwi–Papi and Ago3–Papi interactions. It remains unclear how Papi is maintained in an RNA-free state before the Papi–PIWI association. Also, it remains unclear how piRISC upon its formation is displaced from Papi, and how piRISC avoids re-association with Papi. Conformational change of piRISC may be involved. **b**, *Drosophila* phased piRNA biogenesis involves Zuc–Zuc endonucleolytic cleavage for piRNA 5' and 3' end formation. PIWI-slicer–Nbr exonucleolytic trimming and PIWI-slicer–Zuc

endonucleolytic cleavage produce piRNAs in the ping-pong cycle. The role of *Drosophila* Papi remains under discussion. Its functional homologue(s) (shown as X and Y) may function with Zuc and Nbr. *Bombyx* lacks a gene homologue of *Drosophila* Piwi, so it does not have to accommodate phased piRNA biogenesis. Because of this, Zuc endonuclease might not be used for piRNA 5' end formation. However, PIWI-slicer–Zuc endonucleolytic cleavage produces piRNAs in the ping-pong cycle. We infer that the 3'-to-5' exonuclease cannot trim the 3' end of the intermediate because Papi impedes this reaction. This model shows that *Bombyx* piRNAs are produced in a manner that depends on PIWI-slicer and Zuc. There may also be alternative pathways that have less of an effect on the overall levels of piRNA production.

# The mechanism of eukaryotic CMG helicase activation

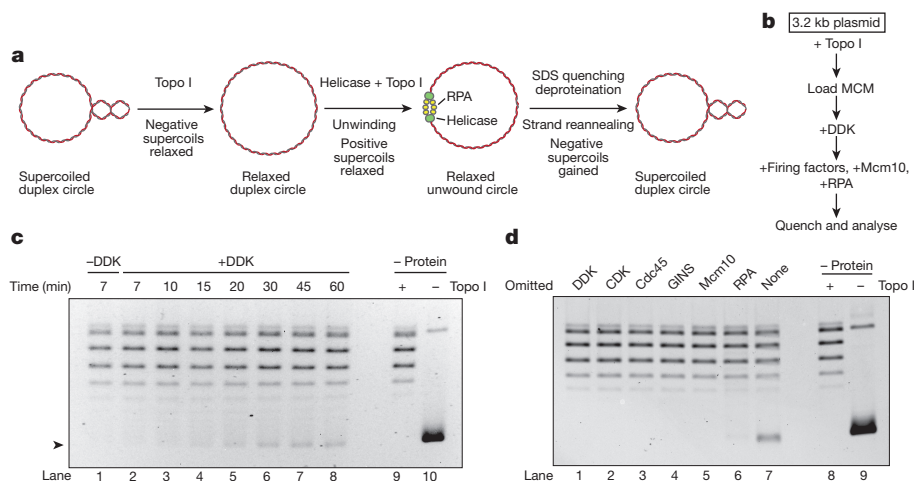
Max E. Douglas<sup>1</sup>, Ferdos Abid Ali<sup>2</sup>, Alessandro Costa<sup>2</sup> & John F. X. Diffley<sup>1</sup>

The initiation of eukaryotic DNA replication occurs in two discrete stages<sup>1</sup>: first, the minichromosome maintenance (MCM) complex assembles as a head-to-head double hexamer that encircles duplex replication origin DNA during G1 phase; then, ‘firing factors’ convert each double hexamer into two active Cdc45–MCM–GINS helicases (CMG) during S phase. This second stage requires separation of the two origin DNA strands and remodelling of the double hexamer so that each MCM hexamer encircles a single DNA strand. Here we show that the MCM complex, which hydrolyses ATP during double-hexamer formation<sup>2,3</sup>, remains stably bound to ADP in the double hexamer. Firing factors trigger ADP release, and subsequent ATP binding promotes stable CMG assembly. CMG assembly is accompanied by initial DNA untwisting and separation of the double hexamer into two discrete but inactive CMG helicases. Mcm10, together with ATP hydrolysis, then triggers further DNA untwisting and helicase activation. After activation, the two CMG helicases translocate in an ‘N terminus-first’ direction, and in doing so pass each other within the origin; this requires that each helicase is bound entirely to single-stranded DNA. Our experiments elucidate the mechanism of eukaryotic replicative helicase activation, which we propose provides a fail-safe mechanism for bidirectional replisome establishment.

Previous studies have focused on the activities of CMG that is preassembled by co-overexpression of individual subunits<sup>4,5</sup>. Here we aimed to understand how CMG is assembled and activated during the initiation of DNA replication, using purified budding yeast proteins<sup>6</sup>. We first used a DNA-topology-based assay<sup>7</sup> (Fig. 1a) to study the unwinding of DNA by CMG. In the presence of topoisomerase I (Topo I), the DNA linking number of a covalently closed circular DNA molecule

decreases by one for each helical turn that is untwisted, allowing changes in DNA unwinding to be inferred quantitatively from changes in DNA supercoiling. We loaded MCM onto a relaxed, circular plasmid in solution, phosphorylated MCM with Dbf4-dependent kinase (DDK), and incubated this with firing factors (defined as Clb5–Cdc28 (hereafter CDK), Cdc45, GINS complex (hereafter GINS), Sld2, Sld3, Sld7, Dpb11, DNA polymerase  $\epsilon$  and Mcm10), replication protein A (RPA), and Topo I in the presence of ATP (Fig. 1b). After incubation, DNA was isolated and plasmid topology was examined by native agarose gel electrophoresis. A fraction of the relaxed plasmid DNA became supercoiled in a time-dependent manner (Fig. 1c), but not when DDK, CDK, Cdc45, GINS, Mcm10 or RPA were omitted (Fig. 1d). These results show that CMG assembled from the double hexamer can unwind DNA even when uncoupled from DNA synthesis, consistent with previous experiments<sup>6,8</sup>. Mcm10 is required for unwinding (Fig. 1d) but not for CMG formation<sup>1,6</sup>. Mcm10 supported extensive unwinding and DNA synthesis even when added after CMG assembly was finished (Extended Data Fig. 1a–d). Thus, Mcm10 activates CMG helicase in a distinct step after CMG assembly (Extended Data Fig. 1e).

The amount of supercoiling that occurs in the absence of RPA should reflect DNA untwisting that is constrained by the CMG (see Extended Data Fig. 2a). To assess this, we constructed a covalently closed, circular 616-base-pair (bp) radiolabelled DNA molecule, which allows us to quantify even small changes in topoisomer distribution relative to the relaxed ground state ( $\alpha$ ) (Extended Data Fig. 2b). CMG assembly and activation in the absence of RPA shifted a proportion of the four starting topoisomers ( $\alpha+1$ ,  $\alpha$ ,  $\alpha-1$ ,  $\alpha-2$ ) to a new set of supercoils,  $\alpha-3$ ,  $\alpha-4$ ,  $\alpha-5$  and  $\alpha-6$  indicating that each circle had been unwound by 3–4 helical turns (Fig. 2a, lane 2). This is likely to be due to genuine

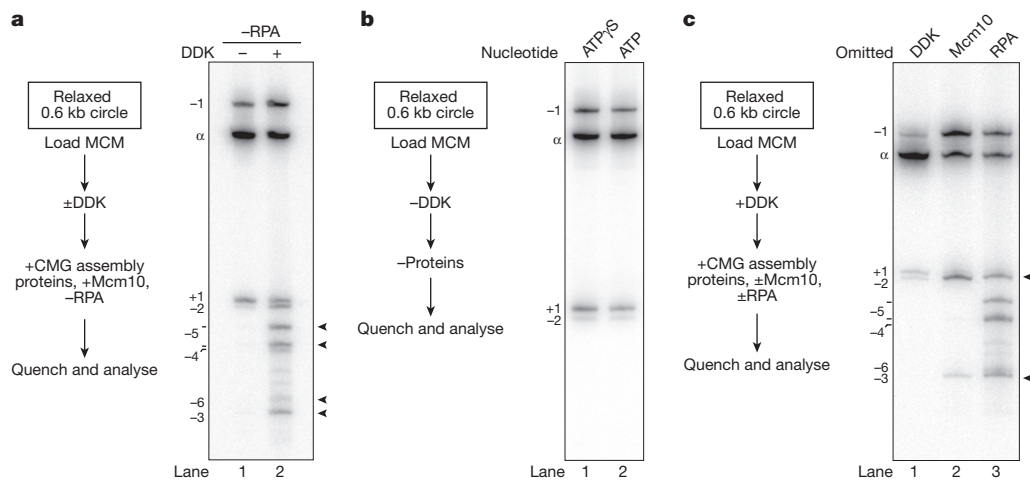


**Figure 1 | Analysis of replicative helicase activation with a DNA unwinding assay.** **a, b**, Outline of the assay. **c**, Time course of unwinding. Purified DNA products were separated on a native agarose gel and stained with ethidium

bromide. No loading or firing factors were added to ‘–protein’ reactions. Arrowhead indicates supercoiled plasmid DNA. **d**, As **c**, with omission of the indicated proteins. Reactions were quenched after 40 min.

<sup>1</sup>Chromosome Replication Laboratory, The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK. <sup>2</sup>Macromolecular Machines Laboratory, The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK.





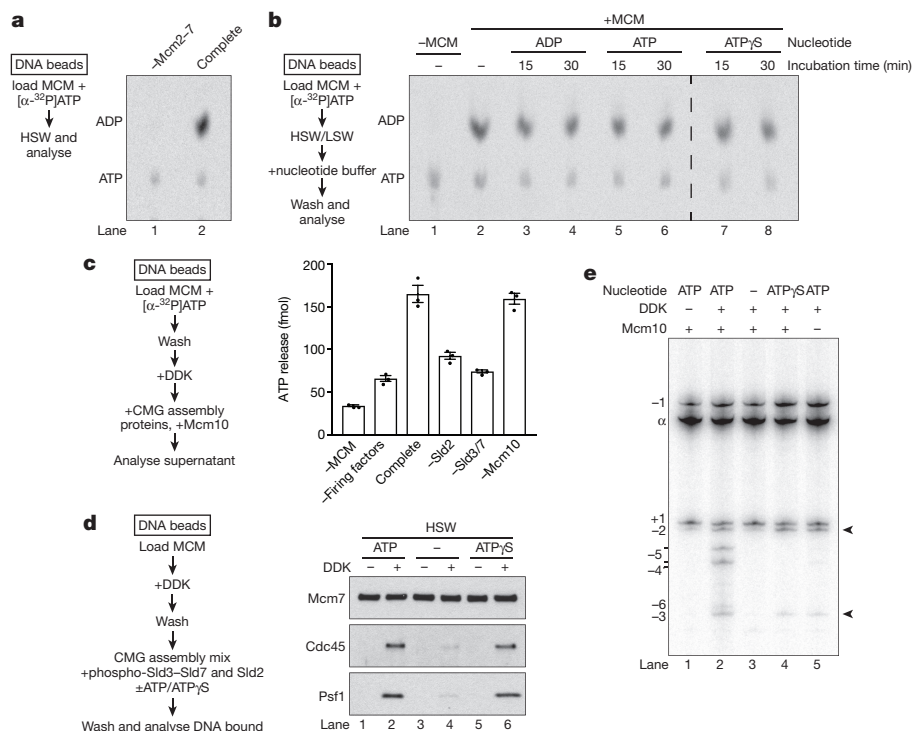
**Figure 2 | Origin unwinding takes place in two steps. a**, Active CMG was assembled on a radiolabelled 616-bp *ARS1* circle in the absence of RPA for 40 min and products were separated on a native bis-polyacrylamide gel. Arrowheads indicate topoisomers observed after CMG assembly and

activation. **b**, As **a**, except all firing factors were omitted. MCM loading does not occur with ATP $\gamma$ S. **c**, As **a**, with omission of the indicated proteins. Arrowheads indicate topoisomers observed after CMG assembly without Mcm10.

unwinding, because thymine residues in DNA became reactive to potassium permanganate (KMnO<sub>4</sub>) across a wide region (Extended Data Fig. 2c). Thus, each of the two activated CMG helicases constrains approximately 1.5–2 helical turns of unwound DNA.

On the basis of the kinked central channel at the interface between hexamers<sup>9,10</sup>, it has been hypothesized that duplex DNA may be distorted by the double hexamer<sup>10</sup>. However, the topoisomer distribution was identical when circles were incubated with loading factors and Topo I in ATP $\gamma$ S (preventing double-hexamer assembly) or ATP

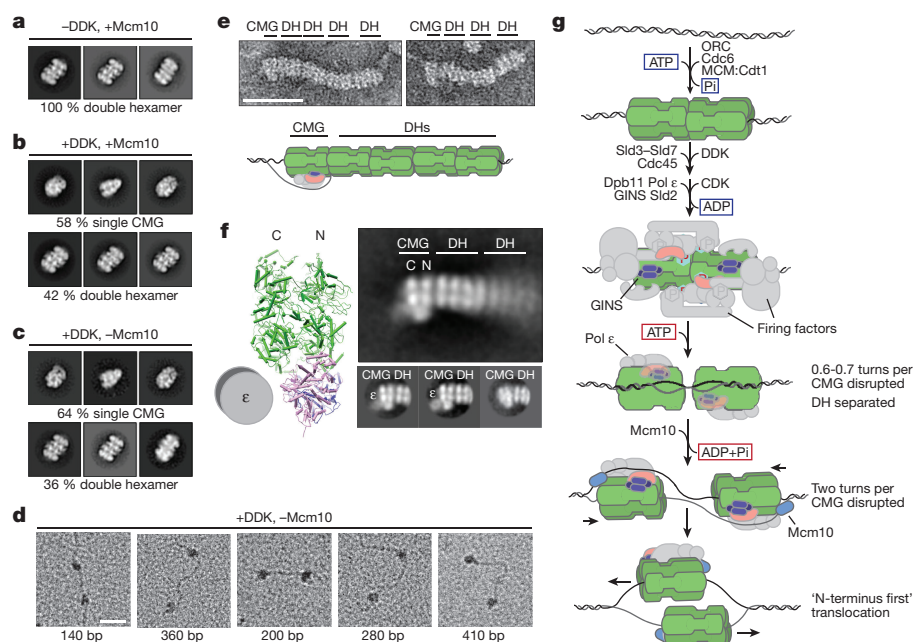
(enabling double-hexamer assembly) (Fig. 2b), indicating that DNA in the double hexamer is not notably untwisted, consistent with recent cryo-electron microscopy (cryo-EM) structures of double hexamer with DNA<sup>11,12</sup>. By contrast, incubation of double hexamer with a full complement of firing factors in the absence of Mcm10 shifted a proportion of circles to more negatively supercoiled topoisomers (Fig. 2c, lane 2:  $\alpha-2$  and  $\alpha-3$ ). Supercoiling required MCM loading (Extended Data Fig. 2d) and all firing factors (Extended Data Fig. 2e), indicating that it takes place during CMG assembly. This shift in supercoiling



**Figure 3 | CMG assembly and activation are coupled to ATP binding and hydrolysis. a**, MCM-loading reactions containing [ $\alpha$ -<sup>32</sup>P]ATP were washed with high-salt buffer (buffer A + NaCl, HSW) and analysed by thin layer chromatography. **b**, As **a**, except washed reactions were incubated with nucleotide as indicated and washed again before analysis. LSW, low-salt buffer. **c**, Double hexamers assembled on bead-immobilized DNA with [ $\alpha$ -<sup>32</sup>P]ATP were used in CMG assembly reactions; the supernatants were analysed by scintillation counting. Error bars, s.e.m.

**d**, Immunoblots of CMG assembly reactions carried out as in Extended Data Fig. 1a, except CDK was omitted, Sic1 was added, Sld2 and Sld3–Sld7 prephosphorylated with CDK were used, and the indicated nucleotide was added. Reactions were quenched 15 min after firing factor addition. **e**, As **d**, except reactions were carried out on soluble *ARS1* circles and analysed as in Fig. 2a. ATP was removed using a spin column after DDK phosphorylation.





**Figure 4 | Structural characterization of replicative helicase activation.** **a–c**, Representative reference-free class averages of helicase activation reactions washed with high-salt buffer (buffer A + KCl). Double hexamer and CMG classes are shown. All particle classes are presented in Extended Data Fig. 4a. **d**, Helicase activation reactions lacking Mcm10 were washed with high-salt buffer and positively stained to visualize DNA. Examples of two CMG-sized particles co-localized with a single DNA fragment are shown. The approximate base-pair distance between particles is indicated.

corresponds to a decrease in linking number of 1.3 (see Methods), indicating that each MCM hexamer untwists around 0.6–0.7 turns of DNA. Origin melting therefore proceeds via two distinct steps: untwisting of 0.6–0.7 turns per MCM hexamer during CMG assembly, and untwisting of approximately one further turn when CMG is activated by Mcm10.

ATP hydrolysis by MCM is required for double-hexamer formation<sup>2,3</sup>, but nothing is known about the downstream roles of ATP in helicase activation. To investigate these roles, we first analysed nucleotide binding and release by the double hexamer. MCM that was loaded in the presence of [ $\alpha$ -<sup>32</sup>P]ATP and washed with high-salt buffer was bound to ADP with only background levels of ATP (Fig. 3a). Without further activation, ADP remained bound and did not exchange with unlabelled ADP, ATP or ATP $\gamma$ S over 30 min (Fig. 3b). To determine whether bound ADP is exchanged during helicase activation, MCM that had been primed with radiolabelled ADP as described above was used as a substrate for CMG assembly. Phosphorylation of MCM by DDK had little effect on the amount of bound ADP (Extended Data Fig. 3a), but ADP was released from the double hexamer into the supernatant of a full helicase activation reaction (Fig. 3c). ADP release was reduced in the absence of Sld2 or Sld3–Sld7, but not in the absence of Mcm10 (Fig. 3c), indicating that ADP release takes place during CMG assembly.

To examine the role of ATP binding and hydrolysis in CMG assembly, double hexamer was loaded onto bead-immobilized DNA, phosphorylated with DDK and washed to remove ATP and DDK. This DDK-phosphorylated double hexamer was then incubated with firing factors, including Sld2 and Sld3–Sld7 which had been phosphorylated by CDK and re-purified to remove ATP and CDK. After low-salt wash, Cdc45 and the GINS (as indicated by the Psf1 subunit) were recruited equally efficiently in a DDK-dependent manner in the presence or absence of nucleotide (Extended Data Fig. 3b, lane 3); however, challenge with high-salt wash showed that CMG assembly did not occur without added nucleotide (Fig. 3d, lanes 2 and 4). ATP $\gamma$ S supported stable CMG formation

(Fig. 3d, lane 6), indicating that CMG assembly does not require ATP hydrolysis. To analyse the effect of nucleotide on DNA untwisting (Fig. 2), reactions were performed as in Fig. 3d, except soluble DNA circles were used as a template and ATP was removed from reactions using spin columns. Whereas no supercoiling occurred in the absence of ATP (Fig. 3e, lane 3), CMG assembly in reactions containing ATP $\gamma$ S and Mcm10 generated the same amount of supercoiling as in reactions containing ATP but lacking Mcm10 (Figs 2c, 3e (lane 5)). Together, these data show that CMG assembly and the initial untwisting of DNA are coupled to ADP release and ATP binding, whereas CMG activation by Mcm10 requires ATP hydrolysis. Mcm10 can stimulate ATP turnover by soluble MCM complex (Extended Data Fig. 3c), and may therefore trigger CMG activation by promoting ATP hydrolysis.

To characterize structural changes that occur during these processes, products assembled on bead-immobilized DNA were washed with high-salt buffer, released from beads by restriction enzyme digestion, and analysed by electron microscopy after negative staining. MCM-containing particles were only detected as double hexamers in reactions lacking DDK, Sld3–Sld7 or Dpb11 (Fig. 4a and Extended Data Fig. 4a, b). By contrast, in a complete reaction, nearly 60% of MCM-containing particles resembled discrete, single CMGs, indicating that approximately 40% of the input double hexamers had been activated (Fig. 4b and Extended Data Fig. 4a). In a reaction containing all firing factors except Mcm10, we observed the same proportion of discrete, single CMGs, after washing with either high- or low-salt buffer (Fig. 4c and Extended Data Figs 4a, c). Separation of the double hexamer therefore takes place during CMG assembly, before CMG is activated. We did not observe double CMGs in these reactions, and pairs of CMG-sized particles that co-localized on the same DNA molecule were separated by up to approximately 400 bp (Fig. 4d and Extended Data Fig. 4d), indicating that inactive CMGs move apart before activation.

CMG translocates in a 3' to 5' direction along the leading strand template<sup>1</sup>, but its orientation at the fork is uncertain. The MCM hexamer is made up of two rings: one formed from the six C-terminal

Scale bar, 50 nm. **e**, Examples of CMGs neighbouring double-hexamer trains (DH) in high-salt-washed reactions on roadblocked DNA. Scale bar, 50 nm. **f**, Annotated reference-free class average from 469 train ends. CMG structure (from Protein Data Bank code 3JC5) is included for reference. 2D classification of train tip particles into several classes (left to right, below) reveals doughnut-shaped polymerase  $\epsilon$  density on a subset of CMGs (left and middle). **g**, Model of eukaryotic replicative helicase assembly and activation.

AAA+ domains and one formed from the N-terminal domains. The hexamers in the double hexamer are linked by their N-terminal domains, and their C-terminal domains are on the outside of the hexamer. Therefore, if the C-terminal MCM ring is at the front of the helicase<sup>13–15</sup>, then the two activated CMGs immediately move away from each other after initiation. If, however, the N-terminal ring is at the front<sup>16</sup>, then the two hexamers must first pass each other during initiation. To investigate these possibilities, we used a 3-kb substrate containing covalent protein roadblocks at the end of each leading strand. We loaded an excess of double hexamers relative to DNA, activated a subset of these with the full set of firing factors, and analysed the products by electron microscopy after negative staining. Figure 4e and Extended Data Fig. 4e show images of multiple, adjacent double hexamers in ‘trains’ with a single CMG at one end. These were not seen if either Mcm10 or the roadblock was omitted (Extended Data Fig. 4f) indicating that active CMG pushes double hexamers, which are free to slide<sup>9</sup>, and the roadblock prevents them from sliding off the DNA end. Comparing 2D averages of the ends of the trains with the structure of the CMG and its 2D projections (Fig. 4f and Supplementary Video 1) indicates that the CMG translocates with the N terminus of MCM in front of the helicase. In agreement with this orientation, a doughnut-shaped density characteristic of polymerase  $\epsilon$ , which binds the C terminus of CMG<sup>17,18</sup>, was observed in a subset of CMGs on the opposite side from the double hexamer (Fig. 4f and Supplementary Video 1).

Our results lead us to propose the model summarized in Fig. 4g. ADP formed during double-hexamer assembly remains stably bound to MCM—recent structural studies suggest that this probably occurs at only a subset of MCM active sites<sup>11,12</sup>. ADP is released in response to firing factors, and subsequent ATP binding by MCM triggers CMG assembly, during which double-hexamer separation takes place. The position of GINS in the CMG is sterically incompatible with the double hexamer<sup>19</sup>, which suggests that these processes occur concomitantly. This step is also accompanied by the first stage of origin melting, when 0.6–0.7 helical turns are unwound per CMG. The earliest steps of origin melting by SV40 large T antigen<sup>7</sup> and *Escherichia coli* DnaA<sup>20</sup> are also triggered by ATP binding, suggesting that this is a conserved feature of replication initiation. The CMG is more than 10 nm in length, but at this stage contains less than 5 nm of ssDNA (6–7 bp fully stretched). Consequently, the lagging template strand cannot yet be entirely excluded from the MCM central channel. Full strand exclusion is required for CMGs to pass one another, indicating that the initial separation of hexamers must take place in the C-terminal direction. CMGs can separate by hundreds of base pairs *in vitro* without Mcm10 (Fig. 4d), but this movement may be restricted *in vivo* by nucleosomes. Moreover, Mcm10 can bind to the double hexamer before firing factor recruitment<sup>21</sup>, which may facilitate immediate activation of CMG. Each active CMG constrains approximately two turns of untwisted DNA, which is long enough (around 15 nm) to be completely excluded from the central channel of MCM. Mcm10 binds avidly to ssDNA, and may therefore play a direct role in this exclusion process<sup>22</sup>. The order and timing of firing factor release is unknown, but Mcm10 has a subsequent role in elongation<sup>8</sup>, suggesting that it may remain bound to the active helicase. Subsequent ‘N terminus first’ crossing of CMGs ensures that all origin DNA will be unwound, and may help to coordinate assembly of the two leading strand replisomes to ensure that this occurs only at origins. Furthermore, the requirement that two helicases can only pass one another when both are bound around ssDNA provides a fail-safe mechanism for bidirectional DNA replication, preventing CMGs from escaping the origin until both helicases are active. The ability of active CMG to push inactive double hexamers ahead of the fork (Fig. 4e, f) may be important for removal of unfired double hexamers from replicated DNA and to rescue stalled forks, but may also necessitate a pathway for double-hexamer removal before termination.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 May; accepted 29 December 2017.

Published online 28 February 2018.

1. Bell, S. P. & Labib, K. Chromosome duplication in *Saccharomyces cerevisiae*. *Genetics* **203**, 1027–1067 (2016).
2. Coster, G., Frigola, J., Beuron, F., Morris, E. P. & Diffley, J. F. X. Origin licensing requires ATP binding and hydrolysis by the MCM replicative helicase. *Mol. Cell* **55**, 666–677 (2014).
3. Kang, S., Warner, M. D. & Bell, S. P. Multiple functions for Mcm2–7 ATPase motifs during replication initiation. *Mol. Cell* **55**, 655–665 (2014).
4. Ilves, I., Petojevic, T., Pesavento, J. J. & Botchan, M. R. Activation of the Mcm2–7 helicase by association with Cdc45 and GINS proteins. *Mol. Cell* **37**, 247–258 (2010).
5. Georgescu, R. E. *et al.* Mechanism of asymmetric polymerase assembly at the eukaryotic replication fork. *Nat. Struct. Mol. Biol.* **21**, 664–670 (2014).
6. Yeeles, J. T., Deegan, T. D., Janska, A., Early, A. & Diffley, J. F. X. Regulated eukaryotic DNA replication origin firing with purified proteins. *Nature* **519**, 431–435 (2015).
7. Dean, F. B. & Hurwitz, J. Simian virus 40 large T antigen untwists DNA at the origin of DNA replication. *J. Biol. Chem.* **266**, 5062–5071 (1991).
8. Lööke, M., Maloney, M. F. & Bell, S. P. Mcm10 regulates DNA replication elongation by stimulating the CMG replicative helicase. *Genes Dev.* **31**, 291–305 (2017).
9. Remus, D. *et al.* Concerted loading of Mcm2–7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139**, 719–730 (2009).
10. Li, N. *et al.* Structure of the eukaryotic MCM complex at 3.8 Å. *Nature* **524**, 186–191 (2015).
11. Noguchi, Y. *et al.* Cryo-EM structure of Mcm2–7 double hexamer on DNA suggests a lagging-strand DNA extrusion model. *Proc. Natl Acad. Sci. USA* **114**, E9529–E9538 (2017).
12. Abid Ali, F. *et al.* Cryo-EM structure of a licensed DNA replication origin. *Nat. Commun.* **8**, 2241 (2017).
13. McGeoch, A. T., Trakselis, M. A., Laskey, R. A. & Bell, S. D. Organization of the archaeal MCM complex on DNA and implications for the helicase mechanism. *Nat. Struct. Mol. Biol.* **12**, 756–762 (2005).
14. Costa, A. *et al.* DNA binding polarity, dimerization, and ATPase ring remodeling in the CMG helicase of the eukaryotic replisome. *eLife* **3**, e03273 (2014).
15. Froelich, C. A., Kang, S., Epling, L. B., Bell, S. P. & Enemark, E. J. A conserved MCM single-stranded DNA binding element is essential for replication initiation. *eLife* **3**, e01993 (2014).
16. Georgescu, R. *et al.* Structure of eukaryotic CMG helicase at a replication fork and implications to replisome architecture and origin initiation. *Proc. Natl Acad. Sci. USA* **114**, E697–E706 (2017).
17. Sun, J. *et al.* The architecture of a eukaryotic replisome. *Nat. Struct. Mol. Biol.* **22**, 976–982 (2015).
18. Zhou, J. C. *et al.* CMG-Pol  $\epsilon$  dynamics suggests a mechanism for the establishment of leading-strand synthesis in the eukaryotic replisome. *Proc. Natl Acad. Sci. USA* **114**, 4141–4146 (2017).
19. Abid Ali, F. *et al.* Cryo-EM structures of the eukaryotic replicative helicase bound to a translocation substrate. *Nat. Commun.* **7**, 10708 (2016).
20. Duderstadt, K. E., Chuang, K. & Berger, J. M. DNA stretching by bacterial initiators promotes replication origin opening. *Nature* **478**, 209–213 (2011).
21. Douglas, M. E. & Diffley, J. F. X. Recruitment of Mcm10 to sites of replication initiation requires direct binding to the minichromosome maintenance (MCM) complex. *J. Biol. Chem.* **291**, 5879–5888 (2016).
22. Robertson, P. D. *et al.* Domain architecture and biochemical characterization of vertebrate Mcm10. *J. Biol. Chem.* **283**, 3338–3348 (2008).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank K. Labib for anti-Psf1 antibody, G. Kelly (the Francis Crick Institute, Bioinformatics) for help with mathematical modelling, the Francis Crick Institute Fermentation Facility for cell production and L. Collinson, R. Carzaniga (the Francis Crick Institute, Electron Microscopy) and T. Pape (Electron Microscopy Centre, Imperial College) for electron microscopy support. This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001065 and FC001066), the UK Medical Research Council (FC001065 and FC001066), and the Wellcome Trust (FC001065 and FC001066). This work was also funded by a Wellcome Trust Senior Investigator Award (106252/Z/14/Z) and a European Research Council Advanced Grant (669424-CHROMOREP) to J.F.X.D.

**Author Contributions** All authors conceived the electron microscopy experiments; M.E.D. prepared the samples and F.A.A. performed the imaging. M.E.D. and J.F.X.D. conceived all other experiments, which were carried out by M.E.D. M.E.D. and J.F.X.D. wrote the paper with input from F.A.A. and A.C.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.F.X.D. ([john.diffley@crick.ac.uk](mailto:john.diffley@crick.ac.uk)).

**Reviewer Information** *Nature* thanks A. Leschziner and the other anonymous reviewer(s) for their contribution to the peer review of this work.



## METHODS

**Protein purification.** All proteins were purified as described<sup>6</sup>.

**DNA templates.** Bead-immobilized linear and circular DNA templates were as described<sup>6</sup>. All plasmid-based assays used pBS/ARS1WTA, a 3.2-kb plasmid containing *ARS1*<sup>23</sup>. To assemble 616-bp *ARS1* circles, a 610-bp fragment around *ARS1* was PCR-amplified from this vector using oligonucleotides oMD171 and oMD172, which introduce recognition sites for EcoRI at both fragment ends. DNA (1.8 µg) was digested with 200 U EcoRI in 100 µl for 3 h at 37 °C, spin-column purified (Roche), and dephosphorylated with 5 U Antarctic phosphatase for 15 min at 37 °C. Phosphatase was inactivated at 70 °C for 5 min, and DNA ends were phosphorylated with PNK and [ $\gamma$ -<sup>32</sup>P]ATP. PNK was heat-inactivated for 20 min at 65 °C, and the sample was desalted over a G50 spin column (GE) and ligated at a concentration of 180 ng/ml overnight at 10 °C with 20 U/ml T4 DNA ligase. The ligation reaction was concentrated 5–10 fold through a 10-kDa-cutoff spin filter (Millipore), ethanol precipitated and run on a 1 × Tris-borate-EDTA (TBE) 1.5% agarose gel. DNA corresponding to supercoiled 616-bp circles was excised and electroeluted for 1 h in 0.1 × TBE. The sample was ethanol precipitated and resuspended in 1 × Tris-EDTA before use.

For the roadblocked DNA template used in Fig. 4e, f, a 2.8-kb fragment containing *ARS1* was amplified using oligonucleotides oMD203 and oMD204, which introduce a single recognition sequence for the methyltransferase HpaII on each end of the *ARS1* fragment. The PCR product was digested with XhoI and cloned into bluescript ks+ digested with XhoI and SmaI to make pMD142. pMD142 was used as a template for PCR with oMD215, which was biotinylated at the 5' end, and oMD208. Methyltransferase HpaII was purified and coupled to this PCR product as described<sup>24</sup> and the coupled DNA product was immobilized on M280 streptavidin resin essentially as described<sup>6</sup>, except resin was washed twice with 10 mM Tris pH 7.2, 1 mM EDTA and 1 M NaCl, twice with 10 mM Hepes pH 7.6, 1 mM EDTA, 1 M KOAc, and twice with 10 mM Hepes 7.6, 1 mM EDTA, and resuspended in half the starting resin volume with 10 mM Hepes 7.6, 1 mM EDTA.

**Oligonucleotides.** The sequences (5' to 3') of oligonucleotides used in this study are: oMD167, CGGAGGTGTGGAGAC; oMD171, TAGTAGGAATT CAAGCAGGTGGGACAGG; oMD172, TAGTAGGAATTCGCGAAAAGACGATAAATACAAG; oMD203, GGTGTATGCATGCTACTGTTTCTCGAGGTG TGAAAGTGGGGTCTCATCTCAGCATCCGGTACCTCAGCGGTAGTTAT AAGAAAGAGACCGAGTTAG; oMD204, GAGCCTGAATCCTCAGCA TCCGGTACCTCAGCAAGAGTATTGGCGATGACGAAAC; oMD208, CCGAAACAGCTATGACCATG; and oMD215, biotin-CGAAAAACCGTCTA TCAGGGCGATG.

**Assigning relative supercoiling states.** To assign the relative supercoiling state of different 616-bp circle topoisomers, 2 fmol/µl 616-bp DNA circles were incubated at 65 °C for 30 min with 0.25 U/µl Nb.BsrDI enzyme, which specifically recognizes and nicks a single site on the circle. Nb.BsrDI was heat inactivated at 80 °C for 20 min, DNA extracted once with phenol:chloroform:isoamylalcohol (25:24:1) and ethanol precipitated. The DNA pellet was resuspended in 1 × Tris-EDTA, and 3 fmol of the DNA was ligated in the presence of the ethidium bromide concentrations indicated at 10–12 °C overnight with 10 U/µl T4 DNA ligase (NEB). Ligated DNA was phenol:chloroform extracted, ethanol precipitated and the DNA pellet resuspended in 1 × Tris-EDTA before analysis by electrophoresis. Final DNA circles are increasingly negatively supercoiled with increasing ethidium bromide during the ligation step. Topoisomers were therefore assigned relative to the ground state ( $\alpha$ , the most prevalent topoisomer when ethidium bromide was omitted) by tracking the order in which bands peaked as the ethidium bromide concentration increased<sup>25</sup>. The nonlinear relationship between increased negative supercoiling and electrophoretic mobility (for example, compare mobility of topoisomers  $\alpha-3$  and  $\alpha-4$ ) has been seen previously<sup>26</sup> and may reflect extrusion of cruciform DNA, which is favoured as linking number decreases<sup>27</sup>.

**Unwinding assays.** Plasmid DNA (25 fmol) or 616-bp DNA (5 fmol) was relaxed in 25 mM HEPES-KOH pH 7.6, 100 mM K-glutamate, 10 mM magnesium acetate, 0.02% NP-40-S, 5% glycerol, 2 mM DTT, 5 mM ATP (loading buffer), with 20 nM Topo I for 30 min at 30 °C. 5 nM ORC, 50 nM Cdc6 and 100 nM Mcm2–7:Cdt1 were added for 20 min at 30 °C, the reaction was supplemented with 50–100 nM DDK, and incubation continued for a further 30 min at 30 °C. Buffer was added to give a final concentration of 250 mM K-glutamate, 25 mM Hepes, 10 mM Mg-acetate, 0.02% NP-40-S, 8% glycerol, 400 µg/ml BSA, 5 mM ATP, 1 mM DTT and 25 nM Topo I (buffer CMG). A mix of firing factors was assembled immediately before use and added at time 0, to a final concentration of 50 nM Dpb11, 200 nM GINS complex, 50 nM Cdc45, 30 nM Pol  $\epsilon$ , 20 nM Clb5–Cdc28 (CDK), 2.5 nM Mcm10, 30 nM Sld3–Sld7, 55 nM Sld2 (firing factor mix). After 40 min at 25 °C (for plasmid DNA) or 30 °C (for small circles), the reaction was quenched with 13 mM EDTA, 0.3% SDS, 0.1 mg/ml Proteinase K (Merck) (stop mix), and incubated at 42 °C for 20 min. Sample was extracted once with phenol:chloroform:

isoamylalcohol (25:24:1), ethanol precipitated, and the DNA pellet resuspended in 1 × Tris-EDTA for analysis.

**Modified unwinding assays.** The experiment in Fig. 2b was carried out as per unwinding assays, except no DDK was used, and no firing factor mix was added after dilution into buffer CMG. The experiment in Fig. 3e was carried out as per unwinding assays, with the following modifications: ATP concentration was reduced to 1 mM for the loading and DDK-phosphorylation steps. After phosphorylation, reactions were passed over a G50 spin column (GE healthcare) washed four times with 25 mM Hepes 7.6, 5 mM Mg-acetate, 10% (vol/vol) glycerol and 0.02% NP-40-S (buffer A) supplemented with 0.1 M K-glutamate. CDK was excluded from the firing factor mix; prephosphorylated Sld2 was used at a final concentration of 10–15 nM, and prephosphorylated Sld3–Sld7 at 10–20 nM. The prephosphorylation procedure is described below. Sic1 was added to a final concentration of 145 nM.

**Gel electrophoresis.** For plasmid-based unwinding assays, DNA was run on native 1.5% agarose Tris-acetate-EDTA (TAE) gels, at 1.5 V/cm for 16 h. Gels were stained with 0.5 µg/ml ethidium bromide for 1 h at room temperature, and destained with 1 mM Mg-sulfate for 1 h before imaging. For 616-bp circle unwinding assays, DNA was run on native 3.5% bis-polyacrylamide 1 × TBE gels at 4.5 V/cm for 20 h.

**Nucleotide-binding analysis.** MCM loading reactions were carried out on 60 ng of immobilized 2.8-kb fragment of *ARS1*<sup>6</sup> in loading buffer with 500 µM ATP, 0.5 µCi/µl [ $\alpha$ -<sup>32</sup>P]ATP, 37.5 nM ORC, 50 nM Cdc6 and 100 nM Mcm2–7:Cdt1. After 30 min at 30 °C, beads were washed twice with buffer A supplemented with 0.5 M NaCl (buffer A + NaCl) and once with buffer A supplemented with 0.25 M K-glutamate and 2 mM CaCl<sub>2</sub> (buffer A + CaCl<sub>2</sub>). DNA-bound complexes were released by cleavage with buffer A + CaCl<sub>2</sub> supplemented with 60 U/µl micrococcal nuclease (MNase) (NEB) for 5 min at 30 °C. Cleaved samples were spotted onto PEI-cellulose TLC plates (Camlab), which were developed in 0.6 M Na<sub>2</sub>HPO<sub>4</sub>–NaH<sub>2</sub>PO<sub>4</sub> pH 3.5. For the nucleotide competition experiments in Fig. 3b, after the NaCl washes described above, DNA-bound complexes were washed once with buffer A + 0.1 M K-glutamate, and incubated at 30 °C for 15 or 30 min in buffer A + 0.1 M K-glutamate and 5 mM of the appropriate nucleotide. Samples were then washed once with buffer A + CaCl<sub>2</sub> and MNase-cleaved and analysed as above.

**Recruitment assays.** MCM-loading, DDK-phosphorylation and CMG-assembly steps were carried out as described in 'Unwinding assays', with the following modifications: each 20 µl CMG assembly reaction used approximately 60 ng linear DNA or 40 ng plasmid DNA immobilized on M-280 streptavidin magnetic beads<sup>6</sup> (Invitrogen). The concentration of ORC was increased to 37.5 nM. Supernatant was removed after DDK phosphorylation and DNA beads were resuspended in buffer CMG without Topo I. After CMG assembly and activation for 8 min, beads were washed twice with 200 µl buffer A with 0.3 M KCl (buffer A + KCl) or twice with buffer A with 0.25 M K-glutamate (buffer A + 0.25 M K-glutamate). After one further wash with 200 µl buffer A + CaCl<sub>2</sub>, beads were resuspended in MNase, and cleaved for 5 min at 30 °C. Supernatant was supplemented with 1/3 volume of 4 × SDS-loading buffer and heated at 95 °C for 3 min. Proteins were separated through 4–12% bis-Tris-polyacrylamide gels (Bio-Rad) and analysed by immunoblotting.

**Modified recruitment assays.** For the experiment in Fig. 2e, MCM loading and DDK phosphorylation were carried out in parallel on 16 fmol *ARS1* plasmid in solution or randomly biotinylated, and immobilized on Streptavidin M-280 resin (Invitrogen). Buffer CMG was added to four soluble reactions, two of which were immediately used to resuspend resin-immobilized reactions for samples 1 and 2. Firing factor mix without Mcm10 was added to all four samples at time 0. After 6 and 10 min, the two remaining soluble reactions were used to resuspend resin-immobilized reactions for samples 3 and 4, respectively. Eight minutes after addition of the soluble reaction to beads, resin was washed twice with 200 µl buffer A + KCl and once with 200 µl buffer A + CaCl<sub>2</sub>, and DNA was cleaved with MNase for 5 min at 30 °C.

Recruitment assays with prephosphorylated Sld2 and Sld3–Sld7 involved the following modifications: the concentration of ATP was reduced to 1 mM for loading and DDK-phosphorylation steps, after which beads were washed three times with buffer A + 0.25 M K-glutamate. CDK was excluded from the firing factor mix and prephosphorylated Sld2 and Sld3–Sld7 were used at 10–15 nM and 10–20 nM, respectively. Sic1 was added to a final concentration of 145 nM.

For the electron microscopy assays in Fig. 4e, f, 120 ng linear, HpaII-coupled DNA immobilized on M-280 streptavidin magnetic beads was used per 20 µl CMG assembly reaction; Mcm2–7:Cdt1 was used at 200 nM during MCM loading.

**Nucleotide-release analysis.** Loading of MCM was carried out with [ $\alpha$ -<sup>32</sup>P]ATP as described in 'Nucleotide-binding analysis'. After 20 min at 30 °C, DDK was added to 100 nM. After a further 30 min at 30 °C, resin was washed twice with buffer A + NaCl and once with buffer A + 0.1 M K-glutamate, and resuspended in buffer CMG without Topo I. At time 0, firing factor mix (described in 'Unwinding assays')

was added. After 15 min at 30°C, supernatant was collected, supplemented with 12.5 mM EDTA, mixed with 5 ml scintillation fluid and measured in a scintillation counter.

**Protein prephosphorylation.** Immediately before phosphorylation, Flag-tagged Sld3–Sld7 was diluted to 30 nM in 40 mM HEPES-KOH pH 7.6, 8% glycerol, 400 µg/ml BSA, 0.02% NP-40-S, 10 mM Mg-acetate, 2 mM DTT, 5 mM ATP with 310 mM K-glutamate (buffer PP + 310 mM K-glutamate). Sld2 was diluted to 120 nM in buffer PP + 235 mM K-glutamate. Clb5–Cdc28 was added to 10 nM, and reactions incubated for 8 min at 25°C before addition of Sic1 to 220 nM. After 2 min incubation at 25°C, Sld2 mix was diluted 4× in buffer A + 0.5 M KCl. Five microlitres magnetic anti-Flag M2 resin (Sigma), washed with buffer A + 0.5 M KCl, was added, and each sample incubated at 4°C for 30 min with rotation. Resin was washed 5× with 300 µl buffer A + 0.5 M KCl (Sld3–Sld7) or buffer A + 0.35 M KCl (Sld2), and resuspended in 10 µl of the same buffer supplemented with 0.25 mg/ml Flag peptide. After shaking at 4°C for 30 min, supernatant was collected, aliquoted and frozen in liquid nitrogen for storage.

**Electron microscopy sample preparation.** For positive and negative stain, CMG assembly was carried out as described in ‘Recruitment assays’. Reactions were washed twice with 200 µl buffer A + KCl and once with 100 µl 25 mM Hepes, 5 mM Mg-acetate, 250 mM K-glutamate (buffer EM), and DNA-bound complexes were released from beads by restriction enzyme cleavage in 5–10 µl buffer EM supplemented with 0.1 U/µl MseI (NEB) for 10 min at 30°C, giving rise to an average DNA fragment size of 1.5–2 kb. Negative-stain sample preparation was performed on 400-mesh copper grids (Agar Scientific) with floated carbon that had been freshly evaporated onto cleaved mica using a Q150TE coater (Quorum Technologies). Grids were glow-discharged for 30 s at 45 mA (Electron Microscopy Sciences). Three-microlitre drops of sample were applied to the grids and left to incubate for 1 min. Excess sample was blotted away and staining was performed on four separate 70-µl 2% uranyl formate drops by stirring for 5, 10, 15 or 20 s. Excess stain was blotted away and grids were stored before imaging. For positive stain, two-week-old carbon-coated 400-mesh copper grids (Agar Scientific) were glow-discharged for 10 s at 45 mA. 3 µl sample was applied and incubated for 30 s. Half the sample solution was blotted away before staining on a single 75-µl 2% uranyl acetate drop for 30 s. Stain was washed away by stirring the grid on four 75-µl ddH<sub>2</sub>O drops for 5 s each before blotting the grid to dryness.

**Electron microscopy data acquisition.** Data collection of negative-stain grids was performed on a Tecnai LaB6 G<sup>2</sup> Spirit transmission electron microscope (FEI) operating at 120 keV (EM STP, The Francis Crick Institute). Micrographs were collected using a 2K × 2K GATAN Ultrascan 100 camera at a nominal magnification of 30,000 (3.45 Å pixel size) or 21,000 (4.92 Å pixel size, train classes) within a –0.5- to 2.5-µm defocus range. Analysis of positive stain grids was performed on a Tecnai G<sup>2</sup> F20 TWIN electron microscope operating at 200 keV (FEI; Electron Microscopy Centre, Imperial College London) equipped with a Falcon II direct electron detector (FEI). Micrographs were collected at a nominal magnification of 50,000 (2.05 Å pixel size) in a defocus range from –3 to –6 µm.

**Electron microscopy single-particle analysis.** Negative-stain particles were semi-automatically picked using EMAN2<sup>28</sup>, version 2.07 and the rest of the image processing was performed using RELION<sup>29</sup> v1.4. Particles were extracted with a box size of 128 × 128 pixels (except the large train class in Fig. 4e, which was extracted with a box size of 250 × 250 pixels) from CTF-corrected (CTFFIND3<sup>30</sup>) micrographs and subjected to reference-free 2D classification with the –only\_flip\_phases additional argument. Comparisons of 2D classes from different samples was performed using the multi-reference alignment function in IMAGIC<sup>31</sup>.

**Electron microscopy positive-stain image analysis.** CMGs were distinguished from MCM double hexamers on positive-stain micrographs by measuring particle length parallel to the DNA axis using ImageJ. Examples where two particles were associated with the same fragment of DNA were analysed, and 84% of particles found to correspond to CMG (57/68). DNA fragments containing two CMG-sized particles are shown.

**Potassium permanganate footprinting.** CMG assembly and activation on small DNA circles were performed as in ‘Unwinding assays’ with the following modifications. The buffer for DNA relaxation and MCM loading was 25 mM Tris-Cl pH 7.2, 100 mM K-glutamate, 10 mM magnesium acetate, 0.02% NP-40-S, 5 mM ATP. After phosphorylation with DDK, buffer was added to give a final concentration of 250 mM K-glutamate, 25 mM Tris-Cl, 10 mM Mg-acetate, 0.02% NP-40-S, 400 µg/ml BSA, 5 mM ATP and 25 nM Topo I. After 10 min of CMG assembly at 30°C, KMnO<sub>4</sub> was added to 3 mM for 4 min before the reaction was quenched with 1 M beta-mercaptoethanol (Sigma) and stop mix, and DNA processed as described in ‘Unwinding assays’. The DNA pellet was resuspended

in 48 µl 1× CutSmart buffer (NEB), digested with 40 U EcoRI-HF (NEB) for 20 min at 37°C, extracted once with phenol:chloroform:isoamylalcohol (25:24:1), ethanol precipitated, and analysed by primer extension. Primer extension reactions contained <sup>32</sup>P end-labelled primer oMD167 and 70 U/ml Vent (exo-) DNA polymerase (NEB), and were carried out for 26 cycles. Reactions were quenched with stop mix, ethanol precipitated and separated on a denaturing 5% urea–bis–polyacrylamide gel.

**DNA replication assay.** For the experiment in Extended Data Fig. 1d, CMG assembly was carried out for 10 min as described in ‘Recruitment assays’, using a randomly biotinylated 5.6-kb *ARS1* plasmid immobilized on M280 streptavidin resin<sup>6</sup>. After CMG assembly in the presence or absence of Mcm10, resin was washed twice with buffer A + KCl (high-salt wash) or buffer A + 0.25 M K-glutamate (low-salt wash), and once with buffer A + 0.25 M K-glutamate. Beads were resuspended in 25 mM Hepes 7.6, 5 mM MgOAc, 0.02% NP-40S, 125 mM K-glutamate, 2 mM ATP, 1 mM DTT, 200 µM CTP, UTP, GTP, 40 µM each dNTP, 40 nM [α-<sup>32</sup>P]dCTP (Perkin Elmer), 50 nM RPA, 30 nM Polymerase ε, 40 mM Polymerase α, 25 mM Topo I and 5 nM Mcm10 or Mcm10 buffer, and incubated for 45 min at 30°C. Reactions were stopped and processed as described<sup>6</sup>.

**MCM ATPase assay.** Mcm2–7 complex was diluted to 0.5 µM in buffer A + 0.5 M K-glutamate, containing 1 µM Mcm10. ATP was added to 100 µM, including 0.125 µCi/µl [α-<sup>32</sup>P]ATP. After 30 min at 30°C, EDTA was added to 15 mM and sample spotted onto PEI-cellulose TLC plates (Camlab), which were developed in 0.6 M Na<sub>2</sub>HPO<sub>4</sub>/NaH<sub>2</sub>PO<sub>4</sub> pH 3.5.

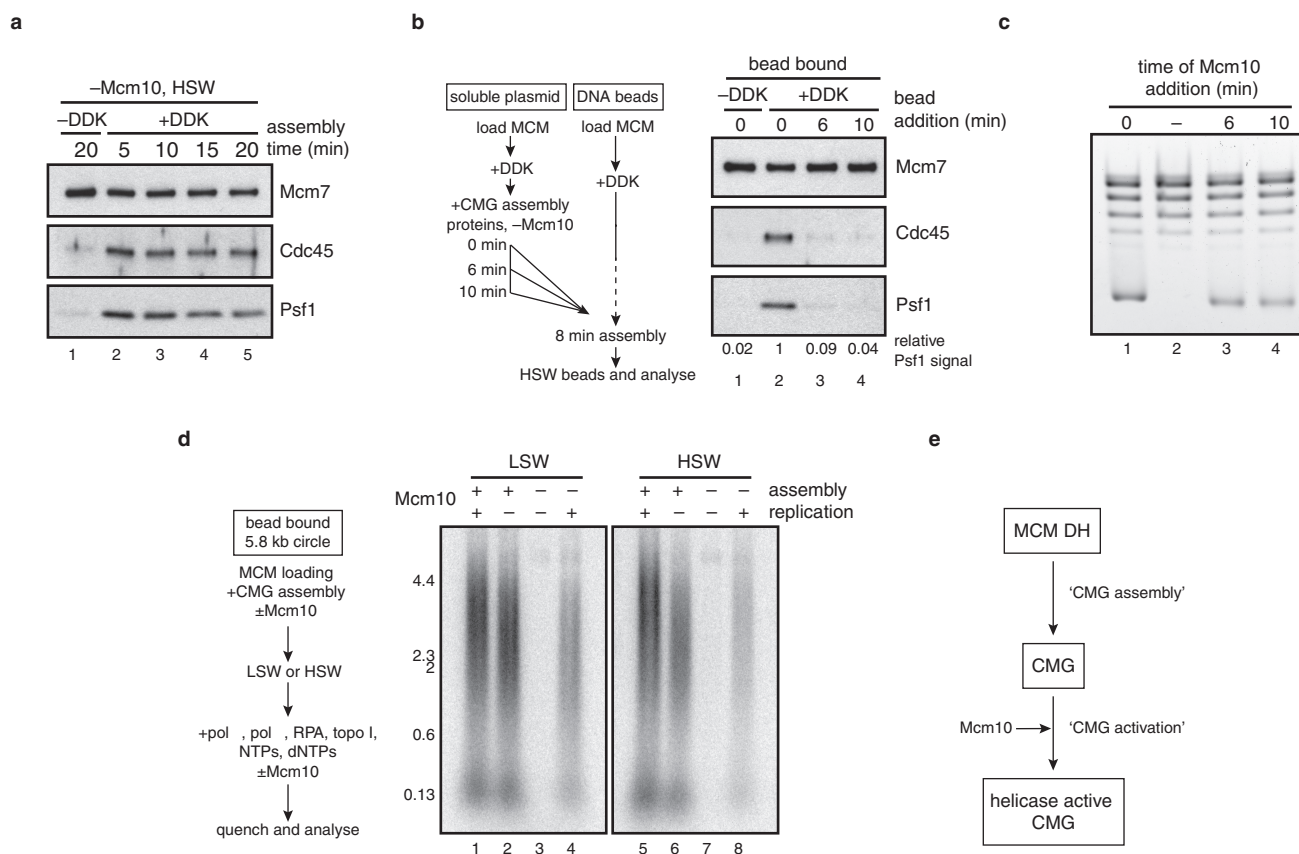
**Estimating linking number shift.** When Mcm10 is omitted from an otherwise complete reaction, CMG is assembled on a fraction of circles, causing a shift in linking number by an unknown amount. If we denote  $Y_k$  as the abundance of a topoisomer  $k$  in the starting distribution of circles without CMG assembly (such as –DDK, lane 1 of Fig. 2c), when CMG is assembled, a proportion ( $a$ ) of circles shift linking number by  $\lambda$ , such that for any  $k$ ,  $a$  of  $k$  gets shifted to another state ( $k-\lambda$ ), while the remaining  $(1-a)$  remains in state  $k$ . We can then model  $X_k$ , the abundance of  $k$  in the –Mcm10 reaction, as  $X_k = Y_{k+\lambda}a + Y_k(1-a)$ , where  $Y_{k+\lambda}a$  is the fraction of topoisomer  $k + \lambda$  that moves into state  $k$  in the –Mcm10 reaction, and  $Y_k(1-a)$  is the amount of topoisomer  $k$  that remains in the –Mcm10 reaction after a fraction has moved to state  $k-\lambda$ . This can be rearranged to  $(X_k - Y_k) = a(Y_{k+\lambda} - Y_k)$  which, given  $\lambda$ , can be solved by linear regression through the origin. Iterating through all possible values of  $\lambda$  on a grid, we choose the one for which residual mean square error was least. To enable fractional offsets to be calculated, we interpolated the original data using a cubic smoothing spline to give us estimated abundances at the resolution of tenths of an integer. A  $\lambda$  value of 1.3 at an efficiency ( $a$ ) of 43% gave the best fit to the measured abundance of topoisomers in the –Mcm10 sample, with an  $R^2$  value of 0.996. This is compared with shifts of 1.1, 1.2, 1.4 and 1.5, which gave  $R^2$  values of 0.9716998, 0.9889401, 0.9888459 and 0.9653362 respectively.

**Statistics and reproducibility.** The experiments in Figs 1d, 2a, 2c, 3a, 3b, 3e and 4a–c were performed at least three times, while the experiments in Figs 1c, 2b, 3d and 4d, e were performed twice. The experiments in Extended Data Figs 1a, 2c–e and 4c were performed at least three times, while the experiments in Extended Data Figs 1b–d, 2b, 3b and 4d–f were performed twice. In Fig. 3c and Extended Data Fig. 3a, c,  $n = 3$  independent experiments.

**Data availability.** The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files.

23. Marahrens, Y. & Stillman, B. A yeast chromosomal origin of DNA replication defined by multiple functional elements. *Science* **255**, 817–823 (1992).
24. Coster, G. & Diffley, J. F. X. Bidirectional eukaryotic DNA replication is established by quasi-symmetrical helicase loading. *Science* **357**, 314–318 (2017).
25. Shore, D. & Baldwin, R. L. Energetics of DNA twisting. II. Topoisomer analysis. *J. Mol. Biol.* **170**, 983–1007 (1983).
26. Zivanovic, Y., Goulet, I. & Prunell, A. Properties of supercoiled DNA in gel electrophoresis. The V-like dependence of mobility on topological constraint. DNA-matrix interactions. *J. Mol. Biol.* **192**, 645–660 (1986).
27. Hsieh, T. S. & Wang, J. C. Thermodynamic properties of superhelical DNAs. *Biochemistry* **14**, 527–535 (1975).
28. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
29. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
30. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).
31. van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116**, 17–24 (1996).

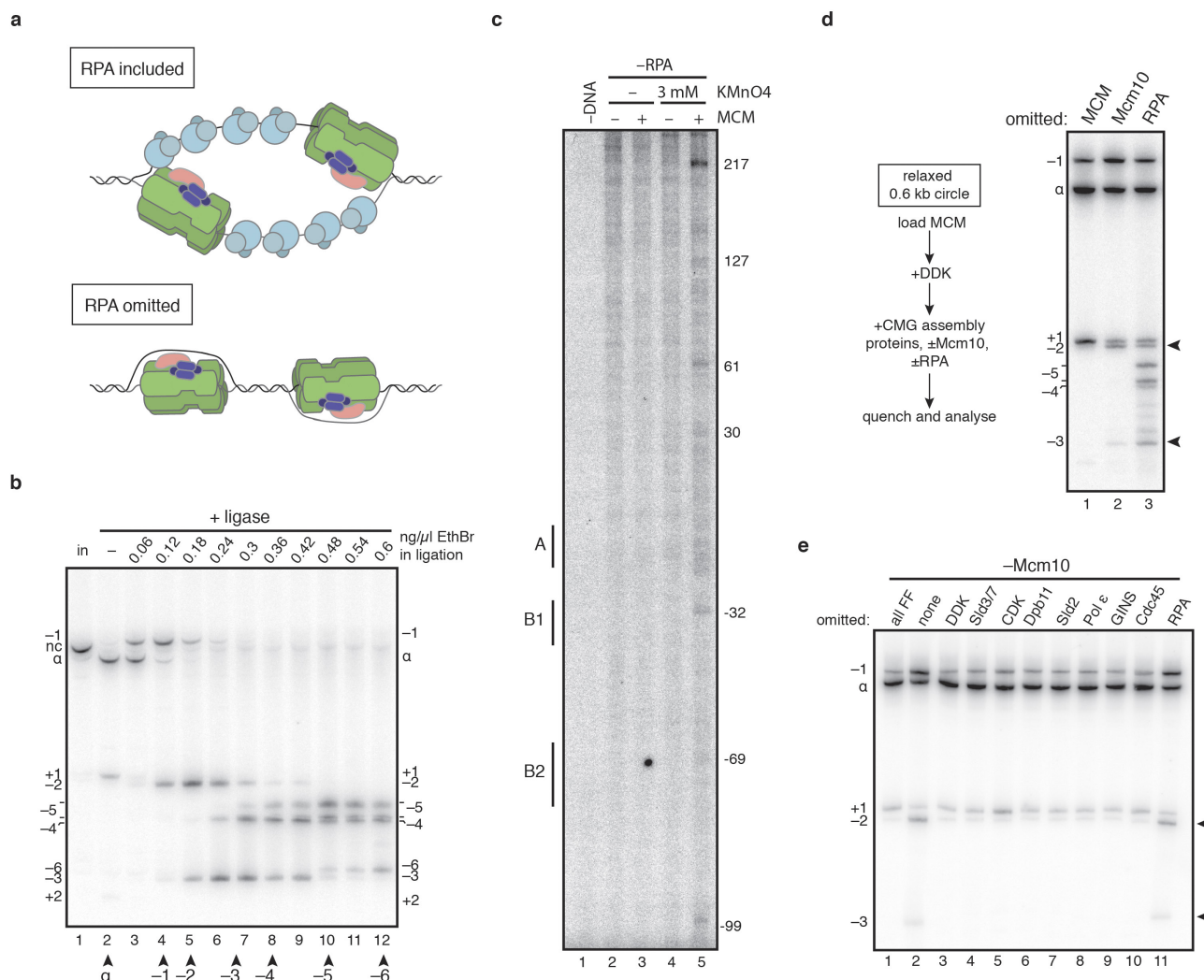




### Extended Data Figure 1 | CMG assembly and activation are separable steps.

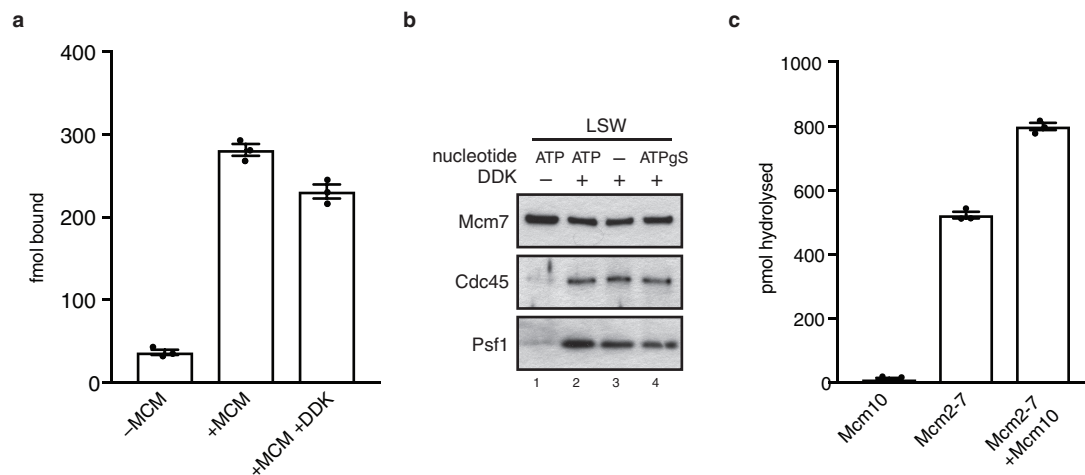
**a**, To determine when CMG assembly saturates, reactions were carried out on bead-immobilized *ARS1* DNA and washed with high-salt buffer (HSW, buffer A + KCl) at the times indicated. The data show that no new CMG assembly takes place after 5 min. **b**, To confirm this, MCMs were loaded in parallel onto a bead-immobilized *ARS1* DNA fragment and a soluble *ARS1* plasmid, and phosphorylated with DDK. A firing factor mix, complete except for Mcm10, was added to the soluble reaction only, which was then added to the bead-immobilized MCMs at the times indicated after firing factor addition to the soluble reaction. After 8 min, beads were washed with high-salt buffer and bound proteins were analysed by immunoblotting. Psf1 signal relative to lane 2 is indicated. The experiment confirms that no CMG assembly takes place more than 5 min

after firing factors have been added. **c**, To test whether Mcm10 can trigger DNA unwinding even after CMG assembly has finished, reactions were set up as in Fig. 1d, except Mcm10 was omitted until the times indicated after firing-factor addition. Mcm10 triggered robust unwinding, even when added more than 5 min after firing factors. Mcm10 can therefore activate preassembled CMG for DNA unwinding. **d**, To test whether Mcm10 can activate preassembled CMG for replication, CMG was assembled on an immobilized *ARS1* plasmid with or without Mcm10. Beads were washed with low- (Buffer A + 0.25 M K-glutamate, LSW) or high-salt buffer, and replication proteins with or without Mcm10 and cofactors, including radiolabelled dCTP, were added. Mcm10 enabled DNA replication even when CMG had been washed to remove excess firing factors. **e**, Schematic outlining the CMG assembly and CMG activation steps described here.



**Extended Data Figure 2 | Characterization of DNA unwinding using small DNA circles.** **a**, Models of DNA unwinding with or without RPA. **b**, To define the relative positions of different topoisomers of radiolabelled 616-bp DNA circles containing *ARS1* (used to analyse small changes in DNA supercoiling in the unwinding assay), nicked circles (nc, lane 1) were ligated closed in the indicated ethidium bromide (EthBr) concentrations. The supercoiling states of different bands of covalently closed DNA were determined relative to the ground state ( $\alpha$ ) by tracking the order in which bands peaked as ethidium bromide concentration increased and DNA was increasingly negatively supercoiled (see Methods for further details). Two bands peaked at the same position for  $\alpha-5$ , and are likely to represent alternative configurations of the  $\alpha-5$  topoisomer. **c**, Primer extension reactions reading the T-rich strand of the ARS-consensus sequence (ACS) of *ARS1* were carried out using 616-bp *ARS1* DNA treated with potassium

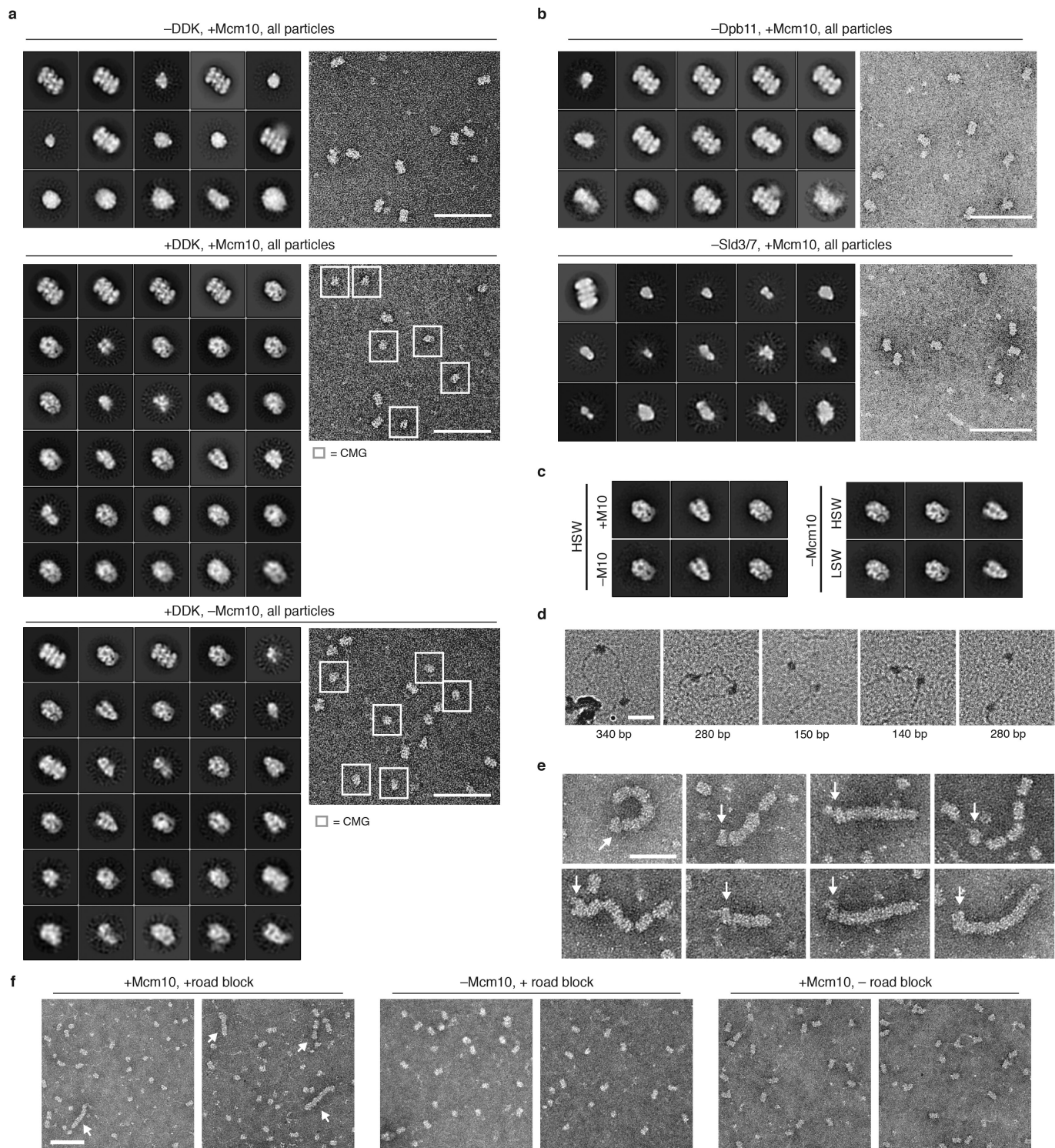
permanganate as indicated after CMG assembly in the absence of RPA. Reactions were separated on 5% sequencing gels, dried and analysed by autoradiography. Base pair numbering is relative to the 5' end of the T-rich strand of the ACS. **d**, As Fig. 2c; lane 1 shows that MCM loading is required for all shifts in topoisomer distribution. Compared with other control samples, such as -DDK, topoisomer distribution was subtly different without MCM; this was not due to loading, which, as shown in Fig. 2b, does not affect topoisomer distribution. **e**, As Fig. 2a, except Mcm10 was omitted from all reactions. No proteins except Topo I were added to the reaction in lane 1 after MCM loading. There was no detectable change in supercoiling relative to when no firing factors (FF) were added (lane 1) when individual firing factors were omitted, suggesting that DNA untwisting in the absence of Mcm10 takes place during CMG assembly.



**Extended Data Figure 3 | Analysis of nucleotide binding and turnover by MCM.** **a**, Double hexamers assembled on bead-immobilized DNA using [ $\alpha$ - $^{32}$ P]ATP were treated with DDK as indicated, and analysed by scintillation counting. Error bars, s.e.m.. **b**, Immunoblots of CMG-

assembly reactions carried out as in Fig. 3d and washed with low-salt buffer. **c**, ATPase assays using [ $\alpha$ - $^{32}$ P]ATP, single-MCM hexamers and Mcm10 as indicated were quantified after thin layer chromatography. Error bars, s.e.m.





**Extended Data Figure 4 | Characterization of replicative helicase activation using electron microscopy.** **a**, Examples of micrographs and complete sets of reference-free class averages of the indicated helicase activation reactions, washed with high-salt buffer (buffer A + KCl). In -DDK, +Mcm10: 7,410 of 23,092 total particles were double hexamers. In +DDK, +Mcm10: 14,668 and 10,492 of 43,320 total particles were CMG and double hexamers, respectively. In +DDK, -Mcm10: 3,984 and 2,226 of 12,920 total particles were CMG and double hexamers, respectively. Classes are positioned with respect to the abundance of

source particles, with the most abundant class in the top left-hand corner, and abundance decreasing from left to right and from top to bottom.

**b**, As **a**, with representative source micrographs. 5,032 of 6,815 and 2,049 of 20,904 particles were double hexamers when Dpb11 or Sld3-Sld7 were omitted, respectively. Scale bar, 100 nm. **c**, Comparison of CMG formed in the indicated conditions. **d**, as Fig. 4d. **e**, As Fig. 4e. Arrows, position of CMG. **f**, Representative crops from micrographs of the indicated samples. Arrows, position of MCM trains. Trains were not observed when either Mcm10 or the protein roadblock was omitted. Scale bar, 100 nm.



# Structure of the D2 dopamine receptor bound to the atypical antipsychotic drug risperidone

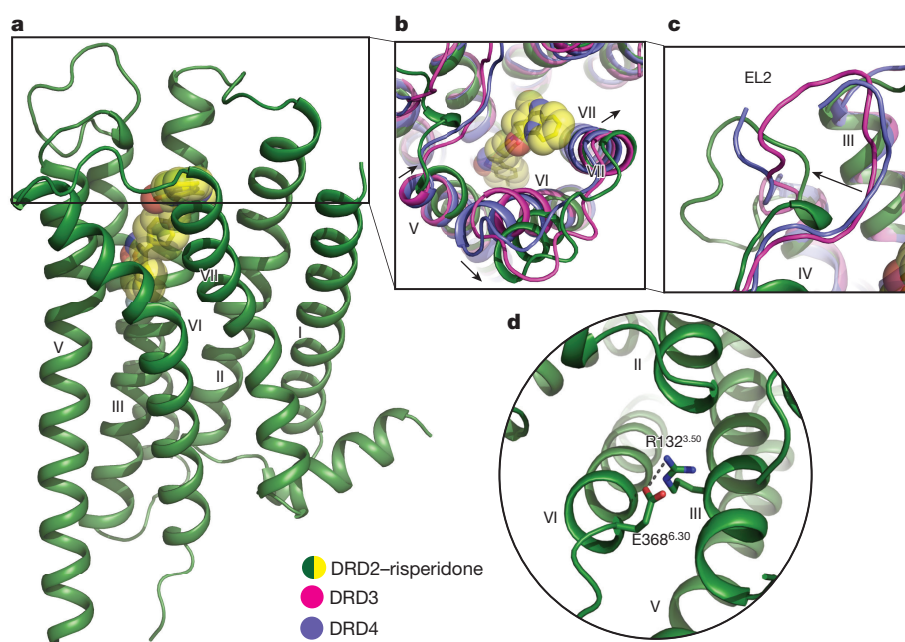
Sheng Wang<sup>1</sup>, Tao Che<sup>1</sup>, Anat Levit<sup>2</sup>, Brian K. Shoichet<sup>2</sup>, Daniel Wacker<sup>1</sup> & Bryan L. Roth<sup>1,3,4</sup>

Dopamine is a neurotransmitter that has been implicated in processes as diverse as reward, addiction, control of coordinated movement, metabolism and hormonal secretion. Correspondingly, dysregulation of the dopaminergic system has been implicated in diseases such as schizophrenia, Parkinson's disease, depression, attention deficit hyperactivity disorder, and nausea and vomiting. The actions of dopamine are mediated by a family of five G-protein-coupled receptors<sup>1</sup>. The D2 dopamine receptor (DRD2) is the primary target for both typical<sup>2</sup> and atypical<sup>3,4</sup> antipsychotic drugs, and for drugs used to treat Parkinson's disease. Unfortunately, many drugs that target DRD2 cause serious and potentially life-threatening side effects due to promiscuous activities against related receptors<sup>4,5</sup>. Accordingly, a molecular understanding of the structure and function of DRD2 could provide a template for the design of safer and more effective medications. Here we report the crystal structure of DRD2 in complex with the widely prescribed atypical antipsychotic drug risperidone. The DRD2–risperidone structure reveals an unexpected mode of antipsychotic drug binding to dopamine receptors, and highlights structural determinants that are essential for the actions of risperidone and related drugs at DRD2.

DRD2 is essential for mediating the actions of antipsychotic drugs<sup>2–4,6</sup> and those of medications used to treat Parkinson's disease, hyperprolactinaemia, and nausea and vomiting, among many other disorders<sup>1,7,8</sup>.

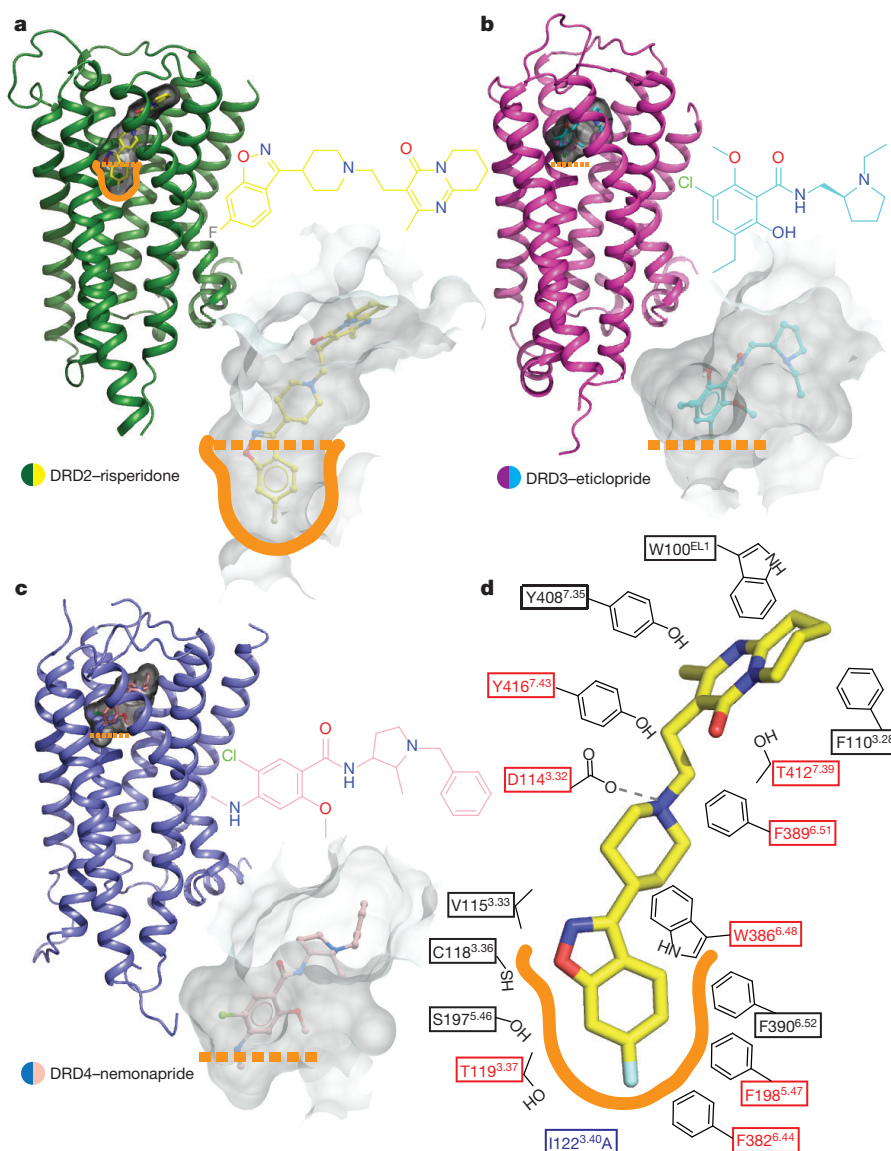
DRD2 has also been implicated in the actions of several drugs of abuse including amphetamines, cocaine and opioids<sup>9</sup>. Although DRD2 was cloned nearly 30 years ago<sup>10–12</sup> and has been subject to extensive pharmacological<sup>13</sup>, mutagenesis<sup>14</sup> and molecular-modelling studies<sup>15</sup>, we lack high resolution structures of DRD2 in complex with ligands, limiting our molecular understanding of its function. A 3.2 Å crystal structure of the related D3 dopamine receptor (DRD3)<sup>16</sup> and 1.95 Å and 2.2 Å structures of the D4 dopamine receptor (DRD4) have been reported<sup>17</sup>. The DRD3 and DRD4 ligand complexes—obtained with the substituted benzamides eticlopride and nemonapride, respectively—revealed distinctive extended binding sites<sup>16,17</sup>. Given the importance of DRD2-targeted drugs, and recent successes in using structures of G-protein-coupled receptors (GPCRs) to guide discovery of new chemical probes and therapeutic leads<sup>18,19</sup>, the structure of DRD2 complexed with non-benzamide ligands will not only clarify the specificity determinants of the family, but will also expand our understanding of how different scaffolds interact with dopamine receptors. We anticipate that the ligand discovery enabled by DRD2 structures will therefore inform both basic and translational neuroscience<sup>20</sup>.

We carried out structural studies using a human DRD2 construct, which included three thermostabilizing mutations (I122<sup>3,40</sup>A, L375<sup>6,37</sup>A and L379<sup>6,41</sup>A; superscript refers to the Ballesteros–Weinstein numbering system for GPCRs<sup>7</sup>) and T4 lysozyme (T4L) fused into intracellular loop 3 (Extended Data Fig. 1a, b and Methods).



**Figure 1 | Structural details of DRD2 and comparison with DRD3 and DRD4.** Dopamine receptor structures are shown aligned to DRD2. Green, DRD2; magenta, DRD3 (PDB code: 3PBL); blue, DRD4 (PDB code: 5WIU). Risperidone (yellow) is shown in sphere representation. **a**, Overall structure of the DRD2–risperidone complex. **b**, **c**, Comparison of the view from the extracellular side. **d**, Cytoplasmic surface showing a salt-bridge interaction (grey dotted line) between R132<sup>3,50</sup> and E368<sup>6,30</sup>.

<sup>1</sup>Department of Pharmacology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7365, USA. <sup>2</sup>Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California 94158-2280, USA. <sup>3</sup>Division of Chemical Biology & Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7360, USA. <sup>4</sup>National Institute of Mental Health Psychoactive Drug Screening Program (NIMH PDSP), School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7365, USA.



**Figure 2 | Comparison of the ligand-binding pocket across the D2-like family receptors.**

**a–c**, Surface representations of the ligand-binding pockets of DRD2 (**a**), DRD3 (**b**, PDB code: 3PBL) and DRD4 (**c**, PDB code 5WIU) are shown in transparent grey. **d**, Schematic representation of risperidone-binding interactions at a 4.0 Å cut-off. Hydrogen bonds are shown as grey dashed lines. The red boxes indicate amino acids that, when mutated, reduce risperidone binding affinity by more than tenfold. The thermo-stabilizing mutation (I122<sup>3.40A</sup>) is shown in blue. The deeper hydrophobic pocket is outlined in orange.

This construct was purified and crystallized in complex with the atypical antipsychotic risperidone. The binding affinities of multiple antipsychotics with this DRD2 construct were similar to those with the wild-type receptor (Extended Data Table 1), suggesting that the alterations that facilitate crystallization do not substantially perturb ligand binding. The crystal structure of the DRD2–risperidone complex was determined at 2.9 Å resolution (Extended Data Table 2 and Extended Data Fig. 1c–h).

Compared with DRD4 (Protein Data Bank (PDB) codes: 5WIU and 5WIV) and DRD3 (PDB code: 3PBL), DRD2 displays substantial structural differences in extracellular loop (EL)1 and EL2, and in the extracellular ends of transmembrane helices (TM)V, TMVI and TMVII (Fig. 1a–c). Unlike in DRD3 and DRD4, the largest extracellular loop of DRD2, EL2, extends away from the top of the receptor core (Fig. 1c). Notably, the highly conserved hydrophobic residue of EL2, which is two residues away from the conserved cysteine of EL2 in all extant aminergic GPCR structures and is represented by Ile184 in DRD2, points towards the receptor core (Extended Data Fig. 2). This residue has been implicated in the on-and-off-rate kinetics and in  $\beta$ -arrestin-biased signalling for some ligands at DRD2 and other receptors<sup>19–21</sup>. However, because of the rearrangement of EL2 and its formation of a small helical turn (residues 182–185) in the

DRD2–risperidone structure (Fig. 1c and Extended Data Fig. 2c), and unlike the analogous residues in some aminergic receptor structures, Ile184 does not directly interact with the ligand (Extended Data Fig. 2a–l). Instead, Ile184 points across the binding pocket to interact with Trp100 in EL1, forming a hydrophobic network near the opening of the binding pocket (Extended Data Fig. 2m). We note that interactions between T4L and EL1 and EL2 in the crystal lattice may further stabilize this conformation (Extended Data Fig. 1c–e), but these weak interactions are unlikely to induce it.

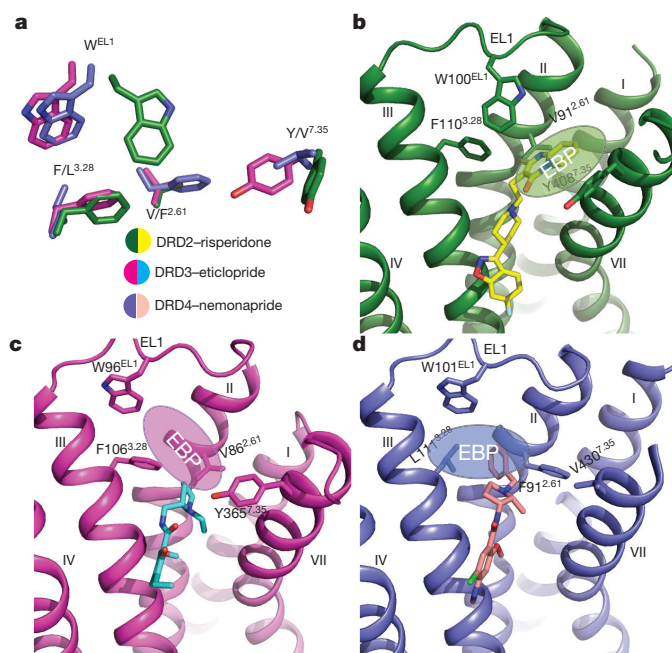
DRD2 also differs from the other two D2-like dopamine receptors in that the extracellular tip of TMV is shifted towards the transmembrane bundle, while the extracellular tips of TMVI and TMVII are approximately 5.8 and 7.3 Å, and 1.4 and 2.1 Å further away from the receptor core, respectively, in comparison to the same regions in DRD3 and DRD4 (Fig. 1b). As in DRD3, an inter-helical hydrogen bond forms between Tyr<sup>7.35</sup> and His<sup>6.55</sup> (Extended Data Fig. 3a–d), which in DRD3 is important for regulating constitutive activity<sup>17</sup>. The side-chain conformations of DRD2, DRD3<sup>16</sup> and DRD4<sup>17</sup> residues Tyr/Val<sup>7.35</sup> and His<sup>6.55</sup> (Extended Data Fig. 3a–c) are also distinct<sup>17</sup>. Specifically, the side chain of Tyr<sup>7.35</sup> in DRD2 is rotated by 52° compared to the one in DRD3 to accommodate risperidone (Extended Data Fig. 3d). Together, these differences may further stabilize the outward movement of TMVI.



Like most antipsychotics, risperidone is a DRD2 inverse agonist<sup>22</sup>, and therefore the DRD2–risperidone complex reflects an inactive state conformation. The most notable difference between active- and inactive-state GPCR structures is the extent to which the cytoplasmic tip of TMVI moves away from the transmembrane helical bundle to accommodate transducer binding<sup>23</sup>. A comparison of DRD2–risperidone with the active and inactive  $\beta_2$  adrenergic receptor ( $\beta_2$ AR) or adenosine  $A_{2A}$  receptor (A2AR) structures reveals no substantial outward movement of the intracellular end of TMVI (Extended Data Fig. 3e, f), a finding consistent with an inactive-state structure. Another important structural feature of GPCR activation is the rearrangement of side chains in the highly conserved microswitches D(E)/RY (TMIII) and NPXXY (TMVII)<sup>23</sup>. Here, Tyr<sup>7.53</sup> from the NPXXY motif and Arg<sup>3.50</sup> from the DRY motif adopt almost identical positions with homologous residues in the  $\beta_2$ AR and A2AR inactive structures (Extended Data Fig. 3g–j). Moreover, a key inactive-state salt-bridge interaction, the ‘ionic lock’ between the conserved Arg<sup>3.50</sup> and Glu<sup>6.30</sup> (refs 24–26) is maintained in the DRD2–risperidone structure (Fig. 1d).

The benzisoxazole risperidone<sup>27</sup> displays a unique mode of dopamine receptor binding in comparison to those of the substituted benzamides eticlopride to DRD3 and nemonapride to DRD4 (Fig. 2). The benzisoxazole moiety of risperidone extends into a deep binding pocket defined by the side chains of TMIII, TMV and TMVI (Fig. 2a, d), and interacts with Cys118<sup>3.36</sup>, Thr119<sup>3.37</sup>, Ser197<sup>5.46</sup>, Phe198<sup>5.47</sup>, Phe382<sup>6.44</sup>, Phe390<sup>6.52</sup> and Trp386<sup>6.48</sup>, which form a subpocket below the orthosteric site (Fig. 2d). Additionally, another hydrophobic pocket above the orthosteric site encloses the tetrahydropyridopyrimidinone moiety of risperidone, whereas Asp114<sup>3.32</sup> forms a salt bridge with the tertiary amine of risperidone (Fig. 2d). Alanine mutagenesis of many of these contact residues reduces the affinity of risperidone binding to DRD2 (Fig. 2d and Extended Data Table 3). In the DRD3 and DRD4 structures, neither eticlopride nor nemonapride engages this deeper hydrophobic pocket (Fig. 2b, c). Importantly, alanine substitutions of the equivalent residues in this deeper hydrophobic pocket do not substantially alter [<sup>3</sup>H]-nemonapride binding affinity for the DRD3 and the DRD4 receptors, except for Trp386<sup>6.48</sup> and Phe390<sup>6.52</sup>, which are large enough that mutagenesis-induced alterations in helical packing alone might explain the observed effects (Extended Data Table 3).

Comparison of the overall ligand-binding pocket of DRD2 with structures of DRD3 and DRD4 revealed marked differences around residues Val/Phe<sup>2.61</sup>, Trp<sup>EL1</sup>, Phe/Leu<sup>3.28</sup> and Tyr/Val<sup>7.35</sup>, which help to define an extended binding pocket (EBP) in DRD2 (Fig. 3a, b). Indeed, previous studies<sup>16,17</sup> on DRD3 and DRD4 revealed a selective EBP in each receptor. The DRD3 EBP is formed by the junction of EL1 and EL2 and the interface of TMII, TMIII and TMVII, and extends towards EL1 (Fig. 3c). By contrast, the DRD4 EBP reaches deep into a cleft between TMII and TMIII, defined by Phe91<sup>2.61</sup> and Leu111<sup>3.28</sup> (Fig. 3d); the structural determination of this DRD4 EBP enabled the structure-based discovery of agonists that are highly specific for this

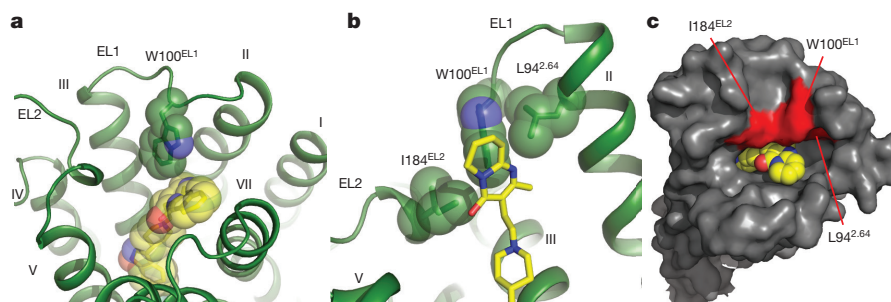


**Figure 3 | Different extended binding pockets revealed across D2-like family receptors.** **a**, The residues of DRD2 (green), DRD3 (pink, PDB code: 3PBL) and DRD4 (blue, PDB code: 5WIU) that define the extended binding pockets (EBPs) are shown as sticks. **b–d**, The distinctive selective EBPs in the D2-like family receptors DRD2 (**b**), DRD3 (**c**) and DRD4 (**d**). Residues and ligands are coloured as in **a**. The position of each EBP is shown as an ellipse.

receptor<sup>17</sup>. Unlike DRD3 or DRD4, the DRD2 EBP extends towards the extracellular part of TMVII, and is formed by EL1 and the junction of TMI, TMII and TMVII (Fig. 3b).

There are four distinctive features of the DRD2 EBP: (1) Compared to the DRD3 structure, part of the EL1 loop is rotated in DRD2, moving the conserved residue Trp<sup>EL1</sup> to the top of the binding pocket (Fig. 3a–c and Extended Data Fig. 4), thereby disrupting what would be the DRD3 EBP (Fig. 3a–c). To our knowledge, this conformation of Trp<sup>EL1</sup> is unique among aminergic receptor structures (Extended Data Fig. 4). (2) The phenylalanine residue is located at 3.28 in DRD2, rather than 2.61 in DRD4, which eliminates the equivalent of the extended pocket of DRD4 in DRD2 (Fig. 3a, b, d). (3) The Tyr408<sup>7.35</sup> side chain rotates towards the His393<sup>6.55</sup> side chain in DRD2, thereby avoiding clashing with risperidone (Extended Data Fig. 3a, d). (4) Finally, an outward movement of the extracellular tip of TMVII (Fig. 1b) makes additional space for the DRD2 EBP.

In comparison with the conformation adopted by risperidone when crystallized in isolation<sup>4</sup>, risperidone’s tetrahydropyridopyrimidinone



**Figure 4 | The hydrophobic ‘patch’ of the DRD2 binding pocket.** **a**, Risperidone (yellow) bound to the orthosteric pocket of DRD2 (green), viewed from the extracellular side. **b**, The W100<sup>EL1</sup> side chain forms

extensive hydrophobic contacts with residues L94<sup>2.64</sup> and I184<sup>EL2</sup>. **c**, The residues L94<sup>2.64</sup>, W100<sup>EL1</sup> and I184<sup>EL2</sup> form a patch (red; residues outside the patch shown in grey) that narrows the binding pocket.

**Table 1 | Risperidone dissociation and association rates with wild-type and mutant DRD2**

Receptor	Residence time, min ( $k_{\text{off}} \pm \text{s.e.m.}, \text{min}^{-1}$ )	$k_{\text{on}} \pm \text{s.e.m.} (\text{M}^{-1} \text{min}^{-1})$	$K_d$ (nM) ( $\rho K_d \pm \text{s.e.m.}$ )
DRD2 wild-type	233 (0.0043 $\pm$ 0.0003)	$1.65 \times 10^6 \pm 1.7 \times 10^5$	2.51 (8.65 $\pm$ 0.21)
DRD2 W100 <sup>EL1</sup> A	28 (0.036 $\pm$ 0.0022) * $P$ =0.007	$5.63 \times 10^6 \pm 3.2 \times 10^5$	6.74 (8.17 $\pm$ 0.04)
DRD2 W100 <sup>EL1</sup> L	23 (0.043 $\pm$ 0.004) * $P$ =0.06	$6.32 \times 10^6 \pm 5.5 \times 10^5$	6.77 (8.17 $\pm$ 0.002)
DRD2 W100 <sup>EL1</sup> F	59 (0.017 $\pm$ 0.002) * $P$ =0.01	$3.12 \times 10^6 \pm 1.8 \times 10^5$	5.30 (8.28 $\pm$ 0.02)
DRD2 L94 <sup>2,64</sup> A	139 (0.0072 $\pm$ 0.0029) NS	$1.43 \times 10^7 \pm 2.3 \times 10^6$	0.48 (9.33 $\pm$ 0.12)
DRD2 I184 <sup>EL2</sup> A	185 (0.0054 $\pm$ 0.002) NS	$9.84 \times 10^6 \pm 1.4 \times 10^6$	0.54 (9.28 $\pm$ 0.10)
DRD2 L94 <sup>2,64</sup> A/I184 <sup>EL2</sup> A	6 (0.16 $\pm$ 0.05) * $P$ =0.005	$2.36 \times 10^7 \pm 7.8 \times 10^6$	7.01 (8.15 $\pm$ 0.02)

Data were acquired by association and dissociation kinetic experiments conducted in parallel at room temperature using [<sup>3</sup>H]-N-methylspiperone (0.8–1.0 nM). Estimates of  $k_{\text{off}}$ ,  $k_{\text{on}}$ , and  $K_d$  were obtained from four independent experiments. Residence time was calculated as  $1/k_{\text{off}}$ . All data are the mean  $\pm$  s.e.m. of four independent assays ( $n=4$  independent experiments). Asterisks indicate statistically significant differences between wild-type and mutant receptors. NS, not significant;  $P$  values are indicated; unpaired two-tailed Student's  $t$ -test.

ring rotates by around 90° in the complex with DRD2 (Extended Data Fig. 5a). This ring interacts with a hydrophobic patch formed by the side chains of Trp100<sup>EL1</sup>, Ile184<sup>EL2</sup>, and Leu94<sup>2,64</sup>. Although the electron density for Leu94<sup>2,64</sup> is weaker than for the other residues, the observed conformation of Trp100<sup>EL1</sup> appears to be stabilized by any rotamer of Leu94<sup>2,64</sup> that would fit the density.

In the DRD2–risperidone structure, the side chain of Trp100<sup>EL1</sup> forms extensive contacts with the tetrahydropyridopyrimidinone ring, wedging it into the DRD2 EBP (Figs. 3b, 4a and Extended Data Fig. 5b). In addition to these hydrophobic contacts between Trp100<sup>EL1</sup> and risperidone, Trp100<sup>EL1</sup> is also stabilized by contacts with Ile184<sup>EL2</sup> and, perhaps, Leu94<sup>2,64</sup>, in spite of the lack of side chain electron density (Fig. 4b and Extended Data Fig. 5c). The observed configuration of risperidone is likely to be driven by the binding pocket of DRD2, and the conservation of key pocket residues such as Trp100<sup>EL1</sup> implies that risperidone could bind other aminergic receptors (for example, 5HT<sub>2A</sub> or the  $\alpha$ 1A adrenergic receptor) in a similar conformation, although further structures will be needed to test this notion.

Molecular docking of risperidone to homology models of DRD2, based on either the DRD3 or DRD4 structures, failed to reproduce the unique pose adopted by risperidone in the complex (Extended Data Fig. 5d–h). Rather, docking placed the ligand higher in the binding site, in a location analogous to that of eticlopride and nemonapride in the DRD3 and DRD4 structures, respectively (Fig. 2b, c). This is a direct consequence of the conformational rearrangements in DRD2 concomitant with accommodating risperidone—mainly movement of TMV, TMVI and TMVII, and the relocation of Trp100<sup>EL1</sup>, which consequently affects the size and shape of the ligand-binding pocket, allowing risperidone to engage a deep binding pose and enter DRD2 EBP. Moreover, the docked conformation of risperidone resembles that of the receptor-free risperidone crystal structure<sup>4</sup>, rather than the conformation adopted in the receptor-bound complex (Extended Data Fig. 5d–h). This is not a problem of conformational sampling on the part of docking—the receptor-free structure is, after all, a low energy structure, and docking captures this—but rather, it reflects the incorrect modelling of Trp100<sup>EL1</sup>, owing to the lack of an analogous configuration in templates used in the modelling. Accordingly, docking does not predict the approximately 90° rotation of the tetrahydropyridopyrimidinone ring of risperidone in the DRD2 complex. The binding pocket of DRD2 and the unusual risperidone conformation that it accommodates are unexpected features of this structure, with implications for our understanding of ligand recognition by this receptor and for the design of new ligands to modulate its activity.

The rearrangement of the extracellular surface and movement of Trp100<sup>EL1</sup> in comparison to the DRD3 and DRD4 structures not only allows it to interact with risperidone, but also forms, together with Ile184<sup>EL2</sup> and Leu94<sup>2,64</sup>, a hydrophobic patch that potentially narrows the binding pocket (Fig. 4b, c). We hypothesized that these residues prevent risperidone from exiting the binding pocket. We found that Trp100<sup>EL1</sup>Phe, Trp100<sup>EL1</sup>Leu and Trp100<sup>EL1</sup>Ala mutations decreased risperidone residence time from 233 min in the wild-type receptor to 59, 23 and 28 min, respectively (Table 1 and Extended Data Fig. 6a–d). Notably, these kinetic effects of the Trp100<sup>EL1</sup>Phe, Trp100<sup>EL1</sup>Leu and Trp100<sup>EL1</sup>Ala mutants on residence time were shared with other

tested antipsychotics, including N-methylspiperone, nemonapride and aripiprazole (Extended Data Fig. 6h–k, o–p and Extended Data Table 4). Similarly, the I184<sup>EL2</sup>A/L94<sup>2,64</sup>A double mutation (Table 1 and Extended Data Fig. 6g) reduced the residence time of risperidone to 6 min, and also reduced the residence times of other antipsychotics (Table 1, Extended Data Fig. 6n, q, r and Extended Data Table 4). In summary, L94<sup>2,64</sup>, Trp100<sup>EL1</sup> and I184<sup>EL2</sup> form hydrophobic contacts that contribute to the slow dissociation of risperidone from DRD2.

Among the most serious side effects of antipsychotics are extrapyramidal symptoms (EPS). A consistent finding in patients with EPS is DRD2 occupancy of more than 80% in the central nervous system, as demonstrated by positron emission tomography (PET)<sup>28</sup>. It has been hypothesized that differential binding kinetics<sup>29,30</sup> and the relatively higher affinity of atypical antipsychotic drugs for 5HT<sub>2A</sub> serotonin receptors<sup>3,4</sup> contribute to the lower incidence of EPS with atypical antipsychotic drugs, such as risperidone, versus typical antipsychotics. We note that Trp100<sup>EL1</sup> regulates both the association and dissociation kinetics of risperidone, and that many of the residues that are essential for risperidone binding to DRD2 are shared with 5HT<sub>2A</sub> serotonin and other biogenic amine receptors. Thus, although our findings do not definitively resolve these hypotheses, they do provide the initial underpinnings for molecularly derived models of the actions of antipsychotic drugs at dopamine and other receptors. Finally, given recent successes in leveraging crystal structures of GPCRs for ligand discovery<sup>17–19</sup>, we anticipate that the DRD2–risperidone complex structure will accelerate the search for novel antipsychotic drugs targeting DRD2.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 9 December 2017; accepted 18 January 2018.**

**Published online 24 January 2018.**

- Missale, C., Nash, S. R., Robinson, S. W., Jaber, M. & Caron, M. G. Dopamine receptors: from structure to function. *Physiol. Rev.* **78**, 189–225 (1998).
- Creese, L., Burt, D. R. & Snyder, S. H. Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs. *Science* **192**, 481–483 (1976).
- Meltzer, H. Y., Matsubara, S. & Lee, J.-C. Classification of typical and atypical antipsychotic drugs on the basis of dopamine D-1, D-2 and serotonin 2 pKi values. *J. Pharmacol. Exp. Ther.* **251**, 238–246 (1989).
- Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3**, 353–359 (2004).
- Roth, B. L. Drugs and valvular heart disease. *N. Engl. J. Med.* **356**, 6–9 (2007).
- Seeman, P. & Lee, T. Antipsychotic drugs: direct correlation between clinical potency and presynaptic action on dopamine neurons. *Science* **188**, 1217–1219 (1975).
- Sibley, D. R. & Monsma, F. J. Jr. Molecular biology of dopamine receptors. *Trends Pharmacol. Sci.* **13**, 61–69 (1992).
- Beaulieu, J. M. & Gainetdinov, R. R. The physiology, signaling, and pharmacology of dopamine receptors. *Pharmacol. Rev.* **63**, 182–217 (2011).
- Volkow, N. D., Fowler, J. S., Wang, G. J., Swanson, J. M. & Telang, F. Dopamine in drug abuse and addiction: results of imaging studies and treatment implications. *Arch. Neurol.* **64**, 1575–1579 (2007).
- Bunzow, J. R. et al. Cloning and expression of a rat D2 dopamine receptor cDNA. *Nature* **336**, 783–787 (1988).
- Grandy, D. K. et al. Cloning of the cDNA and gene for a human D2 dopamine receptor. *Proc. Natl Acad. Sci. USA* **86**, 9762–9766 (1989).



12. Monsma, F. J., Jr, McVittie, L. D., Gerfen, C. R., Mahan, L. C. & Sibley, D. R. Multiple D2 dopamine receptors produced by alternative RNA splicing. *Nature* **342**, 926–929 (1989).
13. Allen, J. A. *et al.* Discovery of  $\beta$ -arrestin-biased dopamine D2 ligands for probing signal transduction pathways essential for antipsychotic efficacy. *Proc. Natl Acad. Sci. USA* **108**, 18488–18493 (2011).
14. Javitch, J. A., Fu, D., Chen, J. & Karlin, A. Mapping the binding-site crevice of the dopamine D2 receptor by the substituted-cysteine accessibility method. *Neuron* **14**, 825–831 (1995).
15. Ballesteros, J. A., Shi, L. & Javitch, J. A. Structural mimicry in G protein-coupled receptors: implications of the high-resolution structure of rhodopsin for structure-function analysis of rhodopsin-like receptors. *Mol. Pharmacol.* **60**, 1–19 (2001).
16. Chien, E. Y. *et al.* Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science* **330**, 1091–1095 (2010).
17. Wang, S. *et al.* D4 dopamine receptor high-resolution structures enable the discovery of selective agonists. *Science* **358**, 381–386 (2017).
18. Manglik, A. *et al.* Structure-based discovery of opioid analgesics with reduced side effects. *Nature* **537**, 185–190 (2016).
19. Wacker, D., Stevens, R. C. & Roth, B. L. How ligands illuminate GPCR molecular pharmacology. *Cell* **170**, 414–427 (2017).
20. McCorvy, J. D. *et al.* Structure-inspired design of  $\beta$ -arrestin-biased ligands for aminergic GPCRs. *Nat. Chem. Biol.* **14**, 126–134 (2018).
21. Free, R. B. *et al.* Discovery and characterization of a G protein-biased agonist that inhibits  $\beta$ -arrestin recruitment to the D2 dopamine receptor. *Mol. Pharmacol.* **86**, 96–105 (2014).
22. Roberts, D. J. & Strange, P. G. Mechanisms of inverse agonist action at D2 dopamine receptors. *Br. J. Pharmacol.* **145**, 34–42 (2005).
23. Rasmussen, S. G. *et al.* Crystal structure of the  $\beta_2$  adrenergic receptor-G<sub>s</sub> protein complex. *Nature* **477**, 549–555 (2011).
24. Shapiro, D. A., Kristiansen, K., Weiner, D. M., Kroeze, W. K. & Roth, B. L. Evidence for a model of agonist-induced activation of 5-HT<sub>2A</sub> serotonin receptors which involves the disruption of a strong ionic interaction between helices 3 and 6. *J. Biol. Chem.* **18**, 11441–11449 (2002).
25. Ballesteros, J. A. *et al.* Activation of the  $\beta_2$ -adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. *J. Biol. Chem.* **276**, 29171–29177 (2001).
26. Palczewski, K. *et al.* Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289**, 739–745 (2000).
27. Janssen, P. A. *et al.* Pharmacology of risperidone (R 64 766), a new antipsychotic with serotonin-S<sub>2</sub> and dopamine-D<sub>2</sub> antagonistic properties. *J. Pharmacol. Exp. Ther.* **244**, 685–693 (1988).
28. Kapur, S., Zipursky, R., Jones, C., Remington, G. & Houle, S. Relationship between dopamine D2 occupancy, clinical response, and side effects: a double-blind PET study of first-episode schizophrenia. *Am. J. Psychiatry* **157**, 514–520 (2000).
29. Kapur, S. & Seeman, P. Does fast dissociation from the dopamine D2 receptor explain the action of atypical antipsychotics?: A new hypothesis. *Am. J. Psychiatry* **158**, 360–369 (2001).
30. Sykes, D. A. *et al.* Extrapyramidal side effects of antipsychotics are linked to their association kinetics at dopamine D2 receptors. *Nat. Commun.* **8**, 763 (2017).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported by NIH Grants R01MH61887, U19MH82441, the NIMH Psychoactive Drug Screening Program Contract and the Michael Hooker Chair for Protein Therapeutics and Translational Proteomics (to B.L.R.) and by R35GM122481 (to B.K.S.). We thank J. Sondek and S. Endo-Streeter for providing independent structure quality control analysis; M. J. Miley and the UNC macromolecular crystallization core for advice and use of their equipment for crystal harvesting and transport, which is supported by the National Cancer Institute under award number P30CA016086; B. E. Krumm for advice on data processing and help with thermostabilization assays; and the staff of GM/CA@APS, which has been funded with Federal funds from the National Cancer Institute (ACB-12002) and the National Institute of General Medical Sciences (AGM-12006). This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science user facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357.

**Author Contributions** S.W. designed experiments, developed the DRD2 construct and purification, expressed, purified and crystallized the receptor, collected diffraction data, solved and refined the structure, analysed the structure, performed radioligand binding and prepared the manuscript. T.C. performed radioligand binding, analysed the data and assisted with preparing the manuscript. A.L. conducted the homology modelling and docking and helped to edit the manuscript. B.K.S. supervised the modelling and docking and helped to prepare the manuscript. D.W. refined and analysed the structure, supervised the structure determination and assisted with preparing the manuscript. B.L.R. supervised the overall project and management and prepared the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.W. ([shengunc@email.unc.edu](mailto:shengunc@email.unc.edu)), D.W. ([dwacker@email.unc.edu](mailto:dwacker@email.unc.edu)) and B.L.R. ([bryan\\_roth@med.unc.edu](mailto:bryan_roth@med.unc.edu)).

**Reviewer Information** *Nature* thanks D. Sibley and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**Protein engineering for structural studies.** To facilitate expression, purification, and crystallography, a human DRD2 (D2 long receptor variant<sup>12</sup>) construct was generated with several modifications. T4L residues 2–161<sup>31</sup> were fused into the third intracellular loop of DRD2 (V223–R361) with truncations of the N-terminal residues 1–34. The DRD2–T4L construct was further modified by introducing three mutations I122<sup>3,40</sup>A, L375<sup>6,37</sup>A and L379<sup>6,41</sup>A, identified by alanine scanning, to improve protein thermostability. In brief, alanine scanning was used to identify thermostabilization mutations (see ‘Radioligand binding assay’ for details; Extended Data Fig. 1a). The chimeric receptor sequences were then subcloned into a modified pFastBac1 vector (Invitrogen), designated pFastBac1-833100, which contained an expression cassette with a haemagglutinin signal sequence followed by a Flag tag, a 10×His tag and a TEV protease recognition site at the N terminus before the receptor sequence.

**Protein expression and purification.** The modified DRD2–T4L protein was expressed in *Spodoptera frugiperda* (Sf9) cells (Expression Systems) using the Bac-to-Bac Baculovirus Expression System (Invitrogen) for 48 h. The insect cells were lysed by repeated washing and centrifugation in hypotonic buffer with low (10 mM HEPES, pH 7.5, 10 mM MgCl<sub>2</sub>, 20 mM KCl and EDTA-free complete protease inhibitor cocktail tablets (Roche)) (once) and high (1.0 M NaCl, 10 mM HEPES, pH 7.5, 10 mM MgCl<sub>2</sub>, 20 mM KCl) salt concentration (three times). The washed membranes were suspended in buffer containing 10 mM HEPES, pH 7.5, 10 mM MgCl<sub>2</sub>, 20 mM KCl, 150 mM NaCl, 20 μM risperidone and EDTA-free complete protease inhibitor cocktail tablets, incubated at room temperature for 1 h and then incubated at 4 °C for 30 min before solubilization. The membranes were then solubilized in 10 mM HEPES, pH 7.5, 150 mM NaCl, 1% (wt/vol) *n*-dodecyl-β-D-maltopyranoside (DDM, Anatrace), 0.2% (wt/vol) cholesteryl hemisuccinate (CHS, Sigma) for 2 h at 4 °C.

The supernatant was isolated by centrifugation at 150,000 *g* for 30 min, followed by incubation in 20 mM buffered imidazole (pH 7.5) and 800 mM NaCl with TALON IMAC resin (Clontech) at 4 °C overnight. The resin was then washed with 10 column volumes of Wash Buffer I (50 mM HEPES, pH 7.5, 800 mM NaCl, 0.1% (wt/vol) DDM, 0.02% (wt/vol) CHS, 20 mM imidazole, 10% (vol/vol) glycerol and 10 μM risperidone), followed by 10 column volumes of Wash Buffer II (25 mM HEPES, pH 7.5, 150 mM NaCl, 0.05% (wt/vol) DDM, 0.01% (wt/vol) CHS, 10% (vol/vol) glycerol and 10 μM risperidone). The protein was then eluted in 3–4 column volumes of Elution Buffer (50 mM HEPES (pH 7.5), 50 μM risperidone, 500 mM NaCl, 10% (vol/vol) glycerol, 0.05% (wt/vol) DDM, 0.01% (wt/vol) CHS, and 250 mM imidazole). A PD MiniTrap G-25 column (GE Healthcare) was used to remove imidazole. The protein was then treated overnight with His-tagged TEV protease and His-tagged PNGase F (NEB) to remove the N-terminal His tag and Flag-tag, and to deglycosylate the receptor. His-tagged TEV protease, His-tagged PNGase F, cleaved His-tag and uncleaved protein were removed from the sample by passing the sample over equilibrated TALON IMAC resin (Clontech). The receptor was then concentrated to 40–50 mg ml<sup>−1</sup> with a 100 kDa molecular mass cut-off Vivaspinn 500 centrifuge concentrator (Sartorius Stedim).

**Lipidic cubic-phase crystallization.** Protein samples of DRD2 in complex with risperidone were reconstituted into the lipidic cubic phase (LCP) by mixing 40% of ~60 mg ml<sup>−1</sup> purified DRD2–risperidone with 60% lipid (10% (wt/wt) cholesterol, 90% (wt/wt) monoolein) using the twin-syringe method<sup>32</sup>. Crystallization trials were performed in glass sandwich plates (Marienfeld) using a handheld dispenser (Art Robbins Instruments), dispensing 50 nl of protein-laden LCP and 1 μl precipitant solution per well. Plates were then incubated at 20 °C. Crystals were obtained from precipitant conditions containing 100 mM Tris/HCl pH 7.8, 230 mM lithium nitrate, 25% PEG400, 4% (±)1,3-butanediol. Crystals grew to maximum size of 40 μm × 40 μm × 10 μm within two weeks and were harvested directly from the LCP matrix using MiTeGen micromount loops and flash frozen in liquid nitrogen.

**Data collection, structure solution and refinement.** Crystallographic diffraction data collection was performed at the 23ID-B and 23ID-D beamlines (GM/CA CAT) at the Advanced Photon Source, Argonne, Illinois using a 10-μm minibeam at a wavelength of 1.0330 Å and a Dectris Eiger-16m or Pilatus 3 6M detector, respectively. The crystals were exposed to 0.5 s of unattenuated beam using 0.5° oscillation per frame. A 97.3% complete data set at 2.90 Å resolution of DRD2–risperidone from 20 crystals was integrated, scaled and merged using HKL3000<sup>33</sup>. Initial phase information was obtained by molecular replacement with the program PHASER<sup>34</sup> using two independent search models: a receptor portion of the DRD4–nemonapride complex (PDB code: 5WIU), and the T4L portion of β2AR–T4L (PDB code: 2RH1) as initial models. Refinement was performed with PHENIX<sup>35</sup> and REFMAC followed by manual examination and rebuilding of the refined coordinates in the program COOT<sup>36</sup> using  $|2F_o - F_c|$ ,  $|F_o - F_c|$ , and omit maps.

**Radioligand-binding assay.** Binding assays were performed using membrane fractions of Sf9 cells expressing the crystallization construct DRD2–T4L

(I122<sup>3,40</sup>A, L375<sup>6,37</sup>A and L379<sup>6,41</sup>A) or membrane preparations of HEK-293T transiently expressing DRD2 (D2 long receptor) and different mutants. HEK-293T cells (ATCC CRL-11268; 59587035; mycoplasma free) were transfected and membrane preparation and radioligand binding assays were set up in 96-well plates as described previously<sup>13</sup>. All binding assays were conducted in standard binding buffer (50 mM Tris, 10 mM MgCl<sub>2</sub>, 0.1 mM EDTA, 0.1% BSA, pH 7.4). For displacement experiments, increasing concentrations of compounds were incubated with membrane and radioligands (0.8–1.0 nM [<sup>3</sup>H]-*N*-methylspiperone or 0.1–0.5 nM [<sup>3</sup>H]-nemonapride) (PerkinElmer) for 2 h at room temperature in the dark. To determine the affinity of nemonapride for DRD2 and different mutants, all assays used at least two concentrations of [<sup>3</sup>H]-nemonapride. The reaction was terminated by rapid vacuum filtration onto chilled 0.3% PEI-soaked GF/A filters followed by three quick washes with cold washing buffer (50 mM Tris HCl, pH 7.4) and quantified as described previously<sup>8</sup>. Results (with or without normalization) were analysed using GraphPad Prism using one-site shift models where indicated.

**Radioligand-based thermostability assay.** Membranes from HEK-293T cells expressing wild-type or mutant human DRD2 were resuspended in binding buffer (50 mM Tris, 10 mM MgCl<sub>2</sub>, 0.1 mM EDTA, 0.1% BSA, pH 7.4). [<sup>3</sup>H]-*N*-methylspiperone was added to the membranes to give a final concentration of 1 nM. The samples were incubated at room temperature for 1 h and then aliquoted into PCR strips. Samples were heated to the desired temperature for exactly 30 min, then cooled down to 25 °C for 30 min. The samples were terminated by rapid vacuum filtration onto chilled 0.3% PEI-soaked GF/A filters followed by three quick washes with cold washing buffer (50 mM Tris HCl, pH 7.4) and quantified as described previously<sup>8</sup>. Results were analysed using GraphPad Prism. Apparent *T<sub>m</sub>* values were derived from sigmoidal dose–response analysis. Results represent the mean ± s.e.m. of three independent experiments.

**Differential-scanning fluorimetry-based thermostability assay.** The thermal stability of purified protein was determined by measuring fluorescence of the thiol-reactive dye BODIPY FL L-cystine (Invitrogen). The standard assay conditions were 20 mM HEPES (pH 7.5), 200 mM NaCl, 0.025% DDM and 10 mM risperidone with protein concentrations of 1 mg ml<sup>−1</sup> and 1 μM BODIPY FL L-cystine. The melting experiments were performed on a StepOnePlus real-time PCR system from Applied Biosystems. The melting curve experiments were conducted (1 °C/min) and recorded using StepOne software from Applied Biosystems. Results were analysed using GraphPad Prism. Apparent *T<sub>m</sub>* values were derived from sigmoidal dose–response analysis. Results represent the mean ± s.e.m. of three independent experiments.

**Ligand association and dissociation radioligand-binding assays.** Binding assays were performed using membrane preparations of HEK-293T cells transiently expressing DRD2 (D2 long receptor) and different mutants at room temperature. Radioligand dissociation and association assays were performed in parallel using the same concentrations of radioligand, membrane preparations and binding buffer (50 mM Tris, 10 mM MgCl<sub>2</sub>, 0.1 mM EDTA, 0.1% BSA, pH 7.4). All assays used at least two concentrations of radioligand (0.5–1.0 nM [<sup>3</sup>H]-*N*-methylspiperone; 0.5–2.0 nM [<sup>3</sup>H]-nemonapride). For dissociation assays, membranes were incubated with radioligand for at least 2 h at room temperature before the addition of 10 μl of 10 μM excess cold ligand to the 200 μl membrane suspension at designated time points. For association experiments, 100 μl of radioligand was added to 100 μl membrane suspensions at designated time points. Time points spanned 1 min to 7 h, depending on experimental conditions and radioligand. For the determination of *k<sub>on</sub>* and *k<sub>off</sub>* for unlabelled risperidone or aripiprazole, membranes containing either wild-type or mutant proteins were incubated with [<sup>3</sup>H]-methylspiperone and several concentrations of risperidone or aripiprazole. Non-specific binding was determined by addition of 10 μM nemonapride. Immediately (at time = 0 min), plates were harvested by vacuum filtration onto 0.3% polyethyleneimine pre-soaked 96-well filter mats (Perkin Elmer) using a 96-well Filtermate harvester, followed by three washes with cold wash buffer (50 mM Tris pH 7.4). Scintillation cocktail (Meltilex, Perkin Elmer) was melted onto dried filters and radioactivity was counted using a Wallac Trilux MicroBeta counter (PerkinElmer). Data were analysed using ‘dissociation-one phase exponential decay’ or ‘association kinetics-two or more concentrations of hot radioligand’ in Graphpad Prism 5.0. The previously determined [<sup>3</sup>H]-*N*-methylspiperone *k<sub>on</sub>* and *k<sub>off</sub>* rates of DRD2 or mutants were used to estimate the *k<sub>on</sub>* and *k<sub>off</sub>* rates of risperidone and aripiprazole using the ‘kinetics of competitive binding’ equation in Graphpad Prism 5.0 as proposed<sup>37</sup>.

**Homology modelling of DRD2.** Sequence alignment for construction of the DRD2 homology models was generated with PROMALS3D<sup>38</sup>, using sequences of human DRD2 (Uniprot accession number: P14416), DRD3 (P35462) and DRD4 (P21917), as well as sequences of available DRD2-family X-ray structures (DRD3, PDB code: 3PBL (chain A)<sup>16</sup> and DRD4, PDB code: 5WIU (chain A)<sup>17</sup>). The

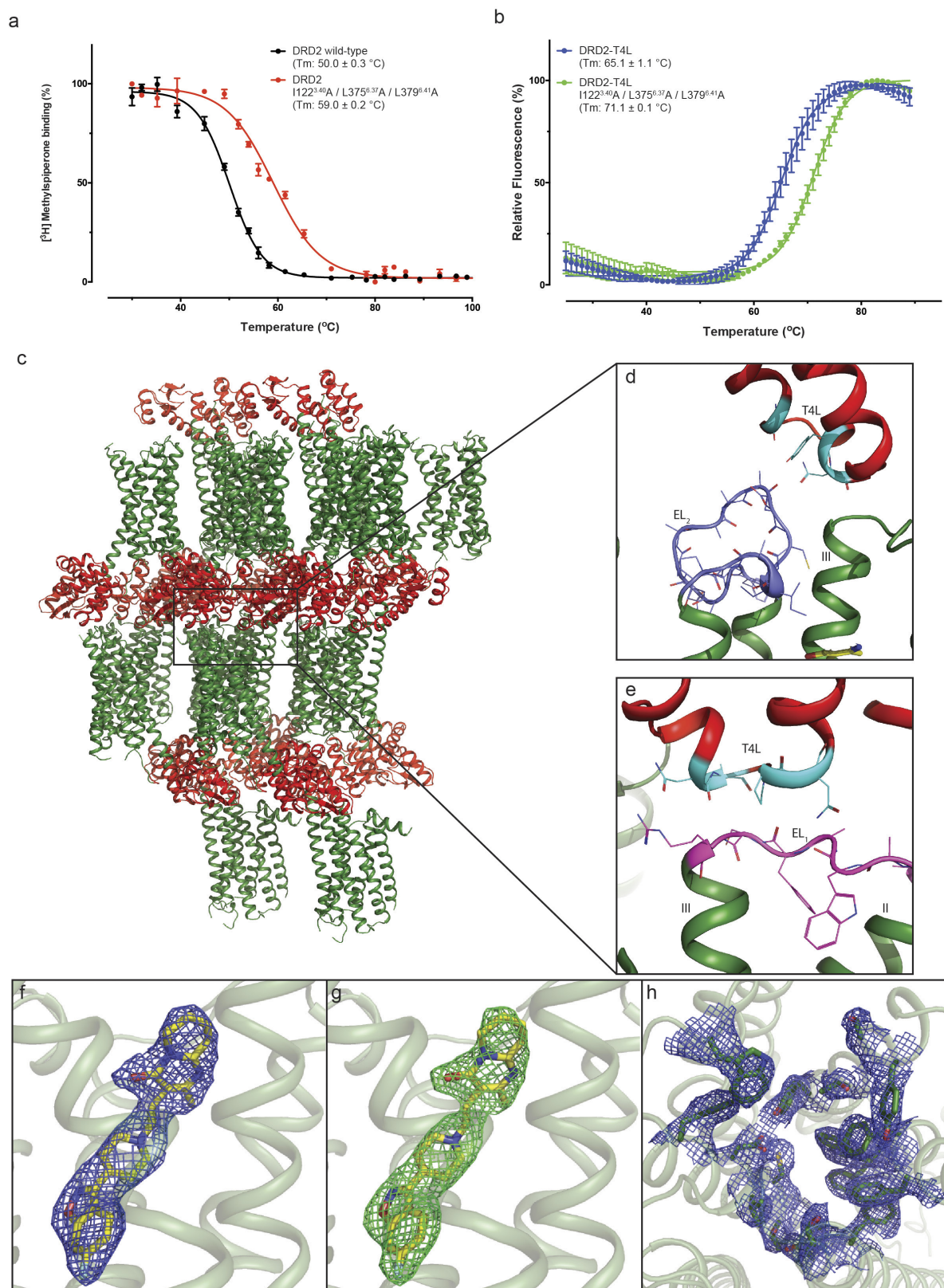
alignment was manually edited to remove the amino and carboxy termini which extended past the template structures, and to remove the engineered T4 lysozyme (PDB code: 3PBL) or apocytochrome b562 RIL (BRIL, PDB code: 5WIU) from the template sequences on DRD4. MODELLER-9v15<sup>39</sup> was then used to generate (1) a total of 1,000 homology models of DRD2, based on the crystal structure of DRD4 in complex with nemonapride as the template, and (2) a set of 500 models based on the crystal structure of DRD3 in complex with eticlopride. We then evaluated the models for their ability to enrich known DRD2 ligands over property-matched decoys through docking to the orthosteric binding site, using DOCK 3.7<sup>40</sup> (as detailed below). While sharing physical properties of known ligands, decoy molecules are topologically distinct and so unlikely to bind the receptor, thus controlling for the enrichment of molecules by physical properties alone. Thirty-two known DRD2 antagonists with molecular weight <420 were extracted from the IUPHAR database<sup>41</sup>, and 1,836 property-matched decoys were generated using the DUD-E server<sup>42</sup>. The models were then ranked on the basis of their adjusted logAUC. The selected best-scoring model in terms of ligand enrichment was further optimized through minimization with the AMBER protein force field and the GAFF ligand force field supplemented with AM1BCC charges<sup>43</sup>.

**Molecular docking of risperidone.** Risperidone was docked to the orthosteric binding site of the DRD2 homology models based on the DRD3 or DRD4 crystal structures using DOCK3.7<sup>40</sup>. DOCK3.7 places pre-generated flexible ligands into the binding site by superimposing atoms of each molecule on matching spheres, representing favourable positions for individual ligand atoms. Forty-five matching spheres were used, based on the pose of the corresponding X-ray ligand (eticlopride or nemonapride) in the template structure. The resulting docked ligand poses were scored by summing the receptor–ligand electrostatics and van der Waals interaction energies, and corrected for context-dependent ligand desolvation. Receptor structures were protonated using Reduce<sup>44</sup>. Partial charges from the united-atom AMBER<sup>43</sup> force field were used for all receptor atoms. Grids that evaluate the different energy terms of the DOCK scoring function were precalculated using AMBER<sup>43</sup> for the van der Waals term, QNIFFT<sup>45,46</sup> (an adaptation of DELPHI) for electrostatics, and ligand desolvation<sup>47</sup>. Ligands were protonated with Marvin (v15.11.23.0, ChemAxon, 2015; <http://www.chemaxon.com>), at pH 7.4. Each protomer was rendered into 3D using Corina<sup>48</sup> (Molecular Networks) and conformationally sampled using Omega<sup>49</sup> (OpenEye Scientific Software). Ligand charges and initial solvation energies were calculated using AMSOL<sup>50,51</sup>.

**Data availability.** Atomic coordinates and structure factor files for the DRD2–Risperidone structure have been deposited in the RCSB Protein Data Bank with identification code 6C38. All other data are available from the corresponding authors upon reasonable request.

31. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into  $\beta_2$ -adrenergic receptor function. *Science* **318**, 1266–1273 (2007).
32. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nat. Protocols* **4**, 706–731 (2009).
33. Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 859–866 (2006).
34. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
35. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
36. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
37. Motulsky, H. J. & Mahan, L. C. The kinetics of competitive radioligand binding predicted by the law of mass action. *Mol. Pharmacol.* **25**, 1–9 (1984).
38. Pei, J. & Grishin, N. V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol. Biol.* **1079**, 263–271 (2014).
39. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics* **47**, 5 6 1–5 6 32 (2014).
40. Coleman, R. G., Carchia, M., Sterling, T., Irwin, J. J. & Shoichet, B. K. Ligand pose and orientational sampling in molecular docking. *PLoS One* **8**, e75992 (2013).
41. Southan, C. *et al.* The IUPHAR/BPS Guide to pharmacology in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res.* **44** (D1), D1054–D1068 (2016).
42. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
43. Case, D. A. *et al.* AMBER 2015. (University of California, 2015).
44. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747 (1999).
45. Gallagher, K. & Sharp, K. Electrostatic contributions to heat capacity changes of DNA–ligand binding. *Biophys. J.* **75**, 769–776 (1998).
46. Sharp, K. A. Polyelectrolyte electrostatics: Salt dependence, entropic, and enthalpic contributions to free energy in the nonlinear Poisson–Boltzmann model. *Biopolymers* **36**, 227–243 (1995).
47. Mysinger, M. M. & Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **50**, 1561–1573 (2010).
48. Sadowski, J., Gasteiger, J. & Klebe, G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **34**, 1000–1008 (1994).
49. Hawkins, P. C., Skillman, A. G., Warren, G. L., Ellingson, B. A. & Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **50**, 572–584 (2010).
50. Chambers, C. C., Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. Model for aqueous solvation based on class IV atomic charges and first solvation shell effects. *J. Phys. Chem.* **100**, 16385–16398 (1996).
51. Li, J., Zhu, T., Cramer, C. J. & Truhlar, D. G. New class IV charge model for extracting accurate partial charges from wave functions. *J. Phys. Chem. A* **102**, 1820–1831 (1998).

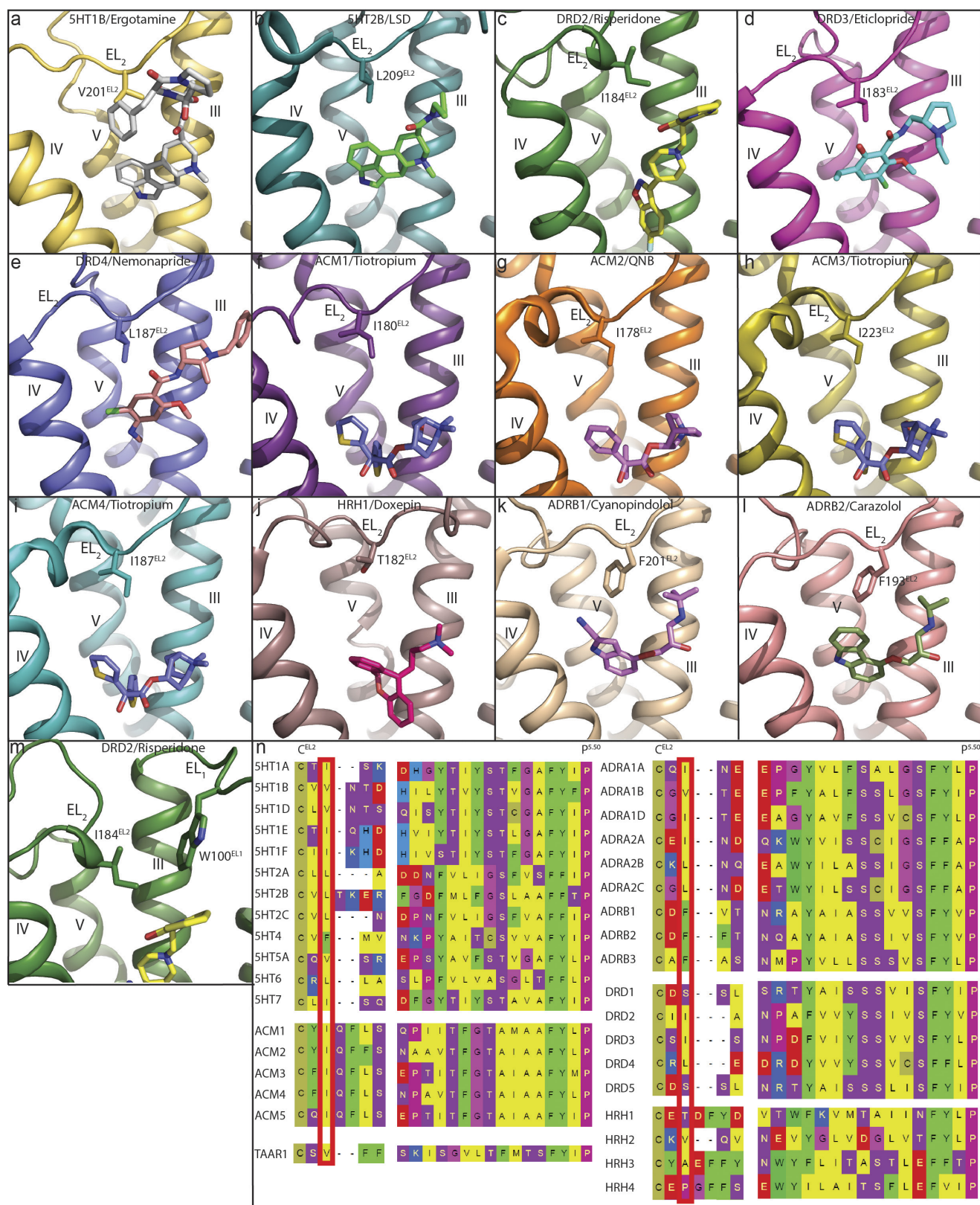




**Extended Data Figure 1 | Thermostability of DRD2 constructs, crystal packing of the DRD2-risperidone complex and representative electron density of the DRD2 structure.** **a**, Membranes containing DRD2 or DRD2 with thermostability mutations were heated for 30 min with 1 nM  $[^3\text{H}]$ -N-methylspiperone and the amount of bound  $[^3\text{H}]$ -ligand was determined. **b**, Purified DRD2-T4L protein (with or without thermostability mutations) was heated with  $10\ \mu\text{M}$  risperidone and  $1\ \mu\text{M}$  BODIPY FL L-cystine dye using a temperature gradient and the amount of dye bound to unfolding protein was determined. Data were analysed by nonlinear regression and apparent  $T_m$  values (transition temperature where 50% of

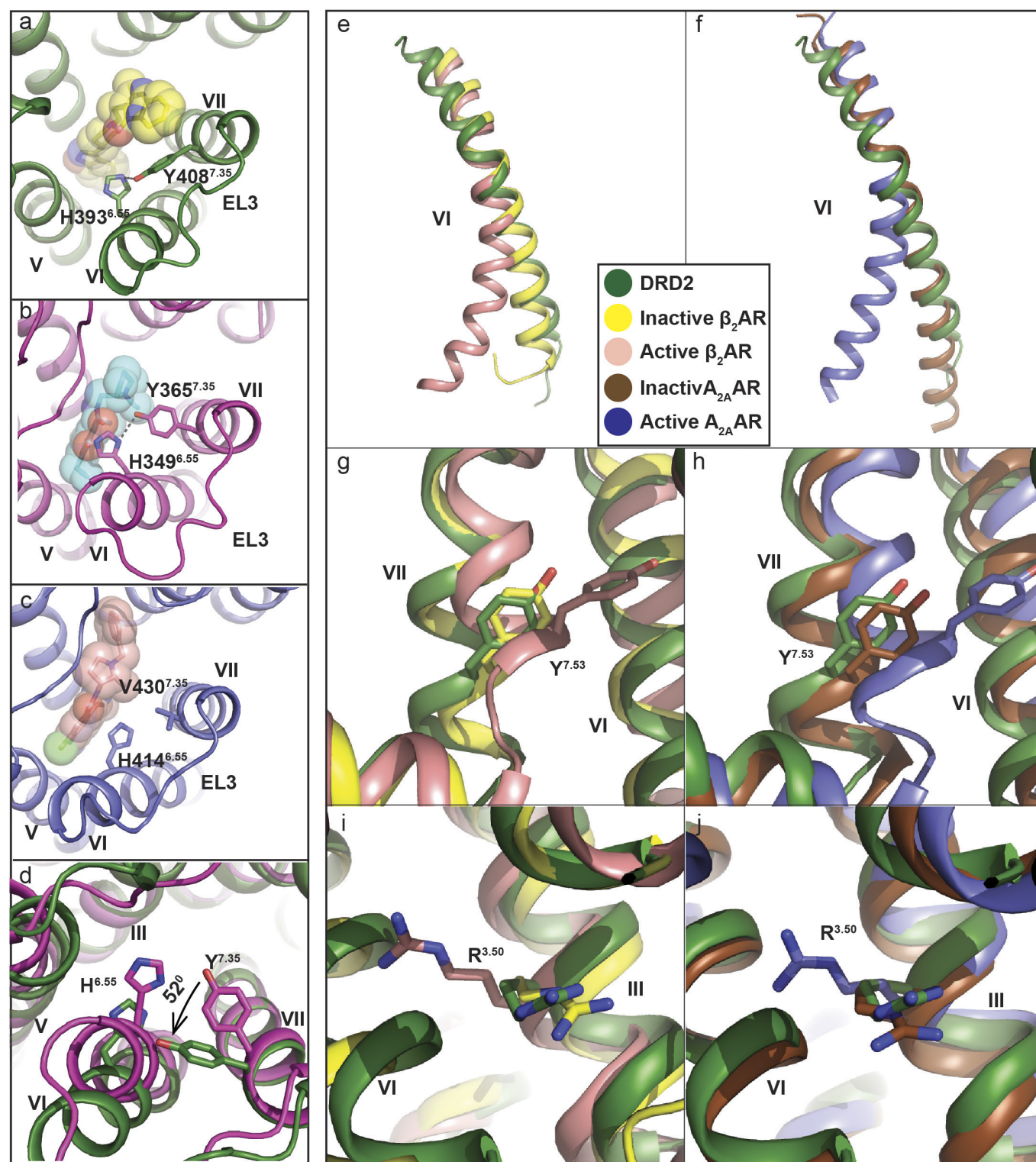
the receptor is inactive) were determined from analysis of the sigmoidal melting curves. All data in **a** and **b** are mean  $\pm$  s.e.m. of three independent assays. **c–e**, Packing of the DRD2-risperidone complex crystallized in the  $P2_12_12_1$  spacegroup. DRD2 is shown in green and the T4L-fusion protein is shown in red, or in cyan where it interacts with DRD2. EL1 and EL2 of DRD2 are shown in magenta and blue, respectively. **f**,  $2F_o - F_c$  electron density map (blue mesh) of risperidone (yellow) contoured at  $1\sigma$ . **g**,  $F_o - F_c$  omit map (green mesh) contoured at  $3.0\sigma$  of risperidone (yellow). **h**,  $2F_o - F_c$  electron density map of DRD2 binding pocket residues (blue mesh) contoured at  $1\sigma$ .





**Extended Data Figure 2 | Conserved hydrophobic residue of EL2 in all available aminergic receptor structures.** In all panels, receptors are shown as cartoons. Ligands and residues are shown as sticks. **a**, 5HT1B (PDB code: 4LAR). **b**, 5HT2B (PDB code: 5TVN). **c**, DRD2 (PDB code: 3PBL). **e**, DRD4 (PDB code: 5WIU). **f**, ACM1 (PDB code: 5CXV). **g**, ACM2 (PDB code: 3UON). **h**, ACM3 (PDB code: 4ADJ). **i**, ACM4 (PDB code: 4DSG). **j**, HRH1 (PDB code: 3RZE). **k**, ADRB1 (PDB code: 2VT4).

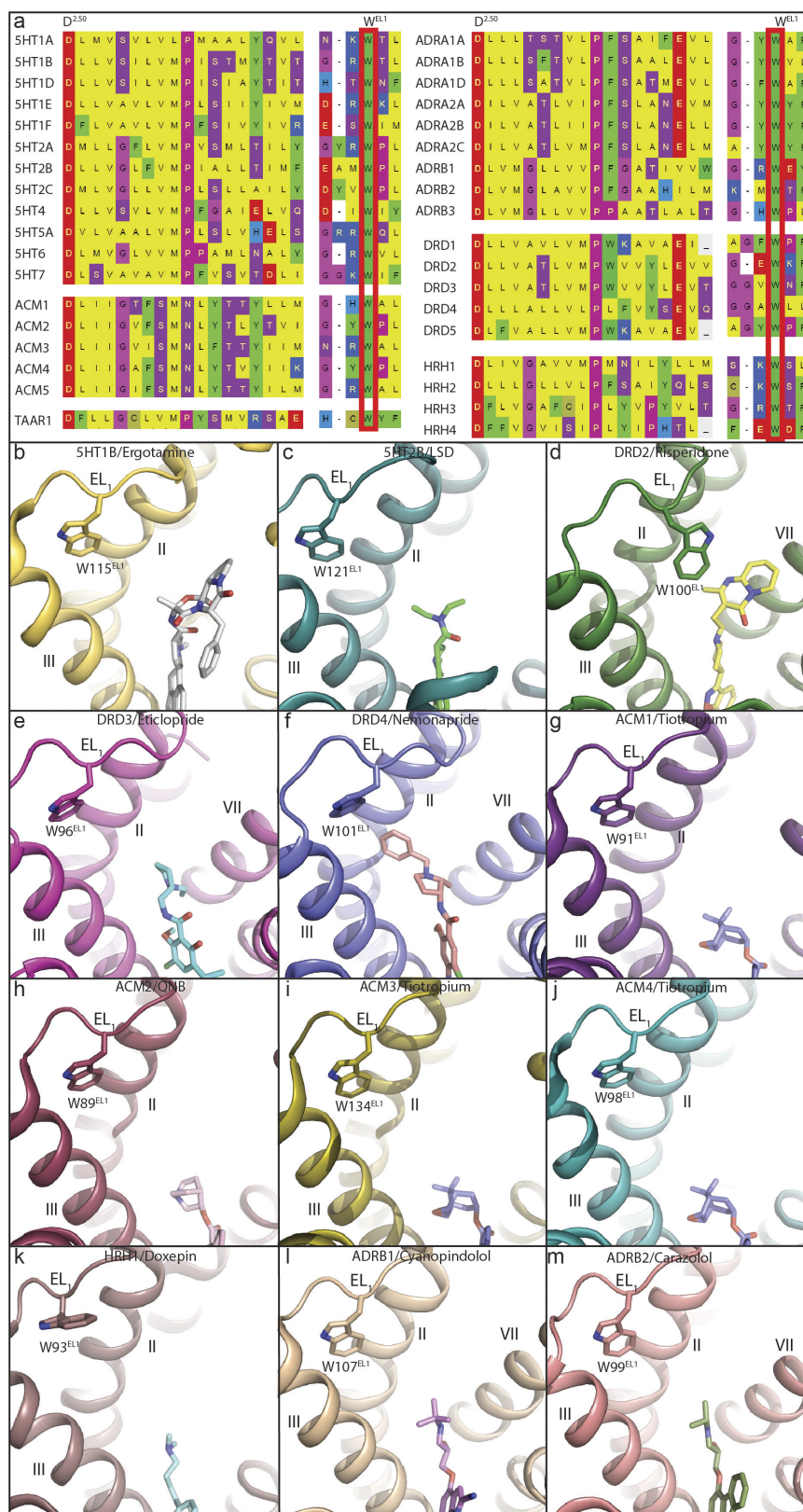
**l**, ADRB2 (PDB code: 2RH1). **m**, DRD2. **n**, Conserved EL2 hydrophobic residues (red box) are located two residues away from a conserved cysteine that forms a disulphide bridge between EL2 and TMIII. Notable exceptions to the presence of a hydrophobic residue are DRD1 and DRD5, which contain a serine, and HRH1 and HRH4, which contain a threonine and proline, respectively.



**Extended Data Figure 3 | Comparison of D2 receptors viewed from the extracellular side, and structural alignment with  $\beta_2$ AR and A<sub>2A</sub>AR reveals an inactive state of DRD2.** a–d, DRD2, green; DRD3, magenta (PDB code: 3PBL); DRD4, blue (PDB code: 5WIU). Risperidone (yellow), eticlopride (cyan) and nemonapride (light pink) are shown as sticks and spheres. Displacements of H<sup>6.55</sup> and Y/V<sup>7.35</sup> are shown at DRD2 (a), DRD3 (b) and DRD4 (c). d, Views from the extracellular side of DRD2 and DRD3. e, f, Superposition of TMVI at DRD2 (green), inactive  $\beta_2$ AR

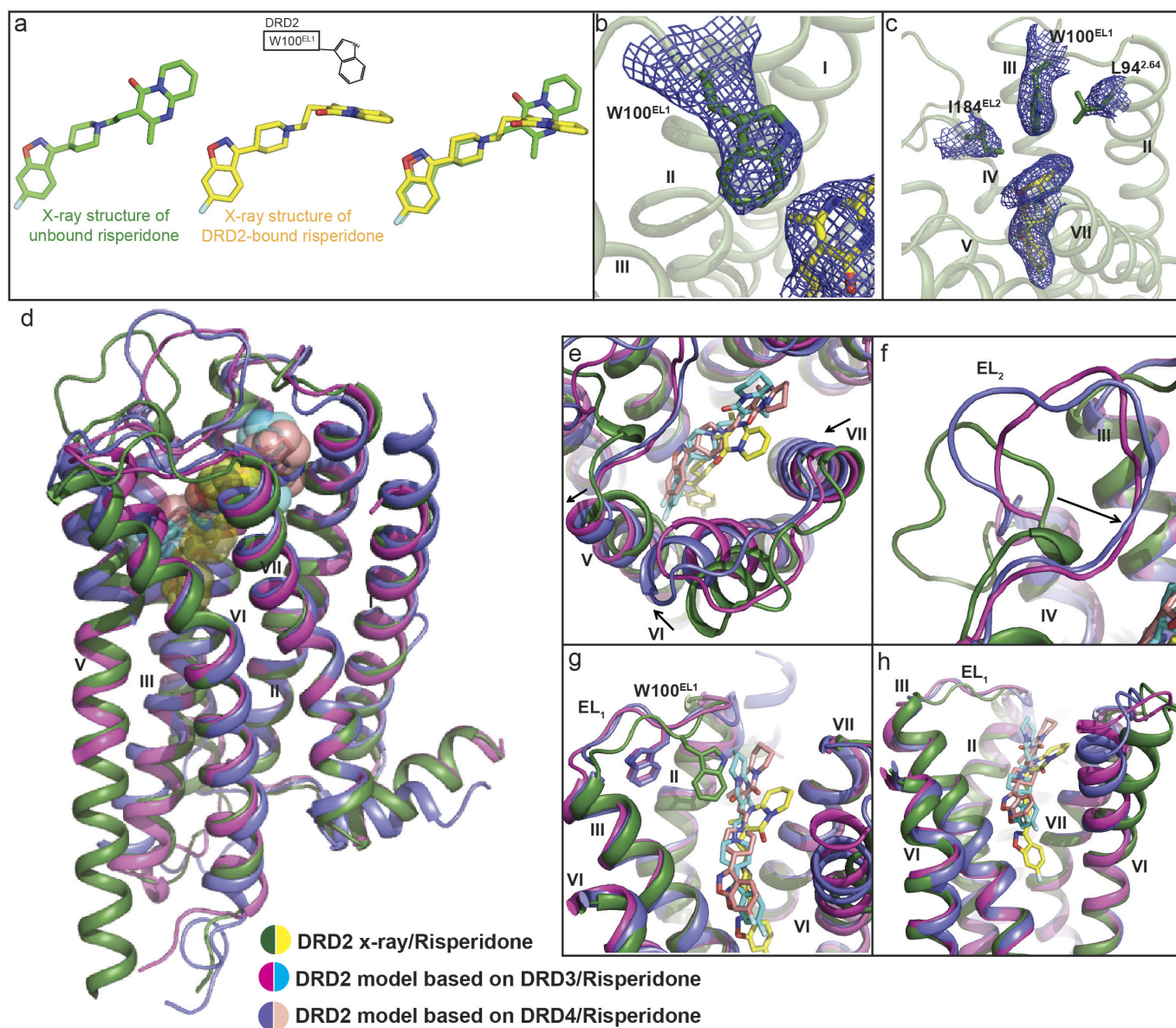
(yellow, PDB code: 2RH1), active  $\beta_2$ AR (light pink, PDB code: 3SN6), inactive A<sub>2A</sub>AR (brown, PDB code: 3REY) and active A<sub>2A</sub>AR (blue, PDB code: 5G53) aligned through helices I–IV. g–j, Cytoplasmic view of alignment between DRD2 and active and inactive  $\beta_2$ AR (g, h) or A<sub>2A</sub>AR (i, j). Rearrangements of two highly conserved residues (Y<sup>7.53</sup> and R<sup>3.50</sup>) within the core of the receptor are shown as sticks. Ligands are omitted for clarity and hydrogen bonds are shown as grey dotted lines.





**Extended Data Figure 4 | Conserved Trp of EL1 in all available aminergic receptor structures shows its unique position in DRD2–risperidone.** Receptors are shown as cartoons. Ligands and residues are shown as sticks. **a**, Conserved Trp residues of EL1 are shown in red boxes. **b**, 5HT1B (PDB code: 4IAR). **c**, 5HT2B (PDB code: 5TVN). **d**, DRD2.

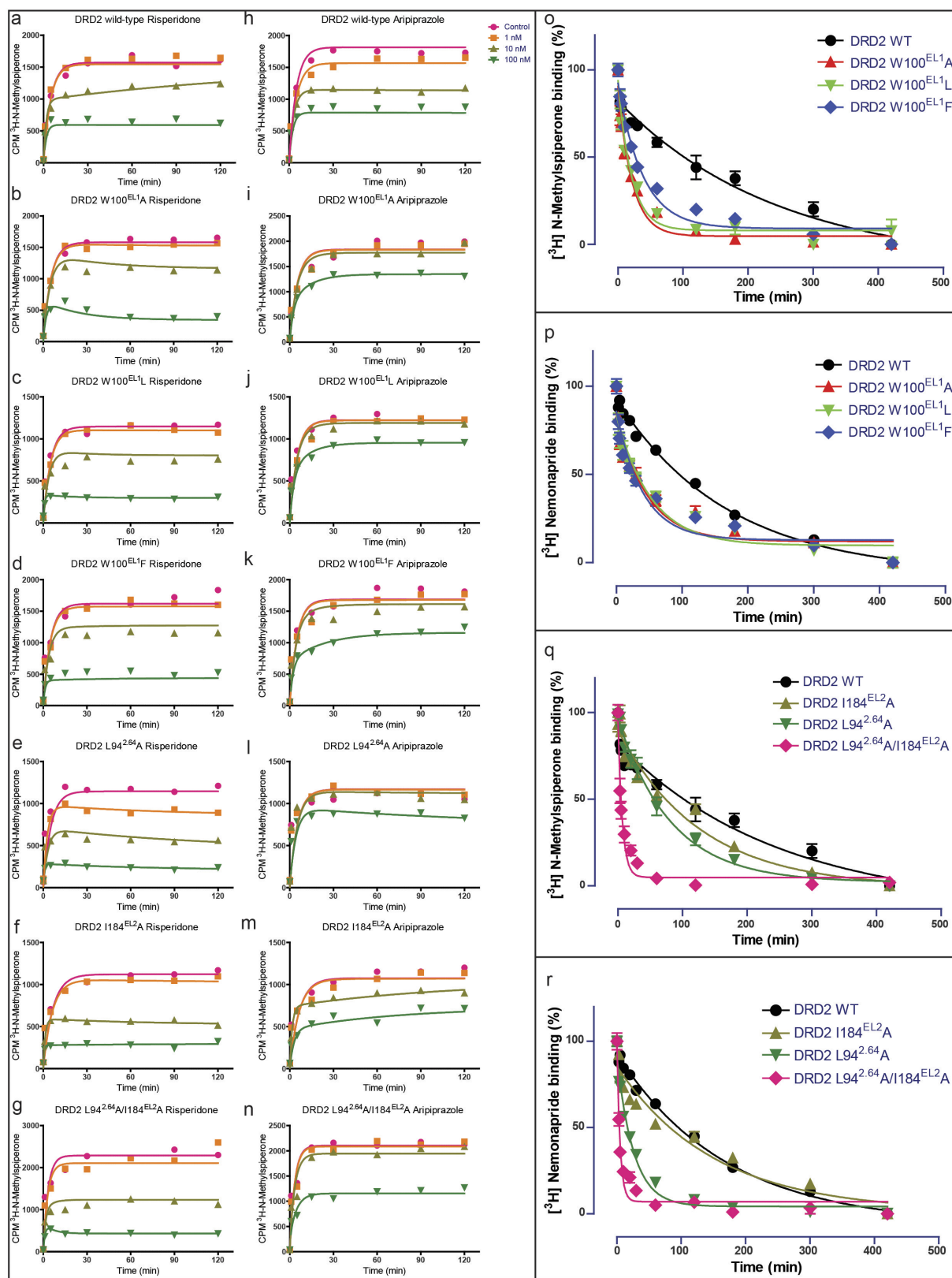
**e**, DRD3 (PDB code: 3PBL). **f**, DRD4 (PDB code: 5WIU). **g**, ACM1 (PDB code: 5CXV). **h**, ACM2 (PDB code: 3UON). **i**, ACM3 (PDB code: 4ADJ). **j**, ACM4 (PDB code: 4DSG). **k**, HRH1 (PDB code: 3RZE). **l**, ADRB1 (PDB code: 2VT4). **m**, ADRB2 (PDB code: 2RH1).



**Extended Data Figure 5 | Risperidone has distinct poses in solution and in complex with DRD2, and comparison of X-ray structure and model of DRD2.** **a**, Trp100<sup>EL1</sup> determines the configuration of the tetrahydropyridopyrimidinone moiety of risperidone. Structure of unbound risperidone shown in green and DRD2-bound risperidone shown in yellow. **b**, Electron density (2F<sub>o</sub>-F<sub>c</sub> maps, blue mesh) for W100<sup>EL1</sup> in the DRD2-risperidone complex (contoured at 1.0σ). **c**, 2F<sub>o</sub>-F<sub>c</sub> electron

density map (blue mesh) of Leu94<sup>2.64</sup>, Trp100<sup>EL1</sup>, Ile184<sup>EL2</sup> and risperidone (yellow) contoured at 0.8σ. **d**, Overall view of DRD2-risperidone X-ray structure and model. **e-h**, Comparison of X-ray structure and model of DRD2. In **d-h**, DRD2 X-ray structure and model are shown as cartoons, with the X-ray structure in green and the model in magenta or blue. Risperidone is shown in the X-ray structure as yellow spheres or sticks, and in the model as cyan or light pink.





**Extended Data Figure 6 | Patch residues of the DRD2 orthosteric pocket impair the dissociation rates of risperidone, aripiprazole, N-methylspiperone and nemonapride.** a–g, Comparison of risperidone dissociation from wild-type DRD2 (a) and W100<sup>EL1</sup>A (b), W100<sup>EL1</sup>L (c), W100<sup>EL1</sup>F (d), L94<sup>2.64</sup>A (e), I184<sup>EL2</sup>A (f) or L94<sup>2.64</sup>A/I184<sup>EL2</sup>A (g) mutants. h–n, Comparison of aripiprazole dissociation from wild-type DRD2 (h) and W100<sup>EL1</sup>A (i), W100<sup>EL1</sup>L (j), W100<sup>EL1</sup>F (k), L94<sup>2.64</sup>A (l), I184<sup>EL2</sup>A (m) or L94<sup>2.64</sup>A/I184<sup>EL2</sup>A (n) mutants. o, p, Comparison of N-methylspiperone

(o) or nemonapride (p) dissociation from wild-type DRD2 and W100<sup>EL1</sup>A, W100<sup>EL1</sup>L or W100<sup>EL1</sup>F mutants ( $n = 3$ ). q, r, Comparison of N-methylspiperone (q) or nemonapride (r) dissociation from wild-type DRD2 and L94<sup>2.64</sup>A, I184<sup>EL2</sup>A or L94<sup>2.64</sup>A/I184<sup>EL2</sup>A mutants. All data are mean  $\pm$  s.e.m. of four independent assays ( $n = 4$  independent experiments). Error bars in o–r denote s.e.m. from four independent assays.

Extended Data Table 1 | Affinities of antipsychotic drugs for thermostabilized mutant and wild-type DRD2

Receptor $K_i$ , nM ( $pK_i \pm SEM$ )	Risperidone	Aripiprazole	N-Methylspiperone	Nemonapride	Bifeprunox
DRD2 wild-type	1.91 ( $8.84 \pm 0.19$ )	6.28 ( $8.21 \pm 0.05$ )	0.04 ( $11.06 \pm 0.18$ )	0.03 ( $11.06 \pm 0.10$ )	1.04 ( $9.52 \pm 0.38$ )
DRD2 I122 <sup>3.40</sup> A / L375 <sup>6.37</sup> A / L379 <sup>6.41</sup> A	1.86 ( $9.10 \pm 0.38$ )	1.25 ( $8.91 \pm 0.04$ )	0.09 ( $11.01 \pm 0.11$ )	0.05 ( $11.03 \pm 0.06$ )	0.24 ( $9.62 \pm 0.02$ )
DRD2-T4L (Sf9) I122 <sup>3.40</sup> A / L375 <sup>6.37</sup> A / L379 <sup>6.41</sup> A	3.13 ( $8.57 \pm 0.18$ )	1.88 ( $8.73 \pm 0.02$ )	0.06 ( $11.02 \pm 0.04$ )	0.09 ( $11.03 \pm 0.33$ )	0.57 ( $9.25 \pm 0.03$ )

Data represent mean  $K_i$  ( $pK_i \pm$  s.e.m.) for competition-binding experiments using [<sup>3</sup>H]-N-methylspiperone (0.8–1.0 nM) as radioligand. All data are the mean  $\pm$  s.e.m of three independent assays ( $n = 3$  independent experiments).

Extended Data Table 2 | Data collection and refinement statistics

Structure	Human DRD2 ( $\Delta N/\Delta ICL3_{T4L}/\Delta C$ )-Risperidone complex
Data Collection	APS, GMCA/CAT 23ID-B/D, 10 $\mu$ m microfocus beam
Crystals	20
Resolution range	30.00 - 2.90 (2.99 - 2.90)
Space group	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>
Unit cell Dimensions a, b, c ( $\text{\AA}$ )	50.98 72.52 151.31
Unique reflections	12826 (889)
Multiplicity	5.5 (2.5)
Completeness (%)	97.3 (86.9)
Mean $I/\sigma(I)$	15.2 (1.0)
$R_{\text{merge}}$ (%)	13.4 (73.8)
$CC_{1/2}$ (%)	99.4 (53.5)
<b>Refinement Statistics</b>	
Reflections used in refinement	12826 (889)
Reflections used for R-free	622 (40)
R-work (%)	22.6 (37.4)
R-free (%)	24.9 (34.1)
<b>Number of Atoms</b>	
DRD2	1948
T4L	1176
Risperidone	30
Lipid and other	82
<b>Overall B-factors (<math>\text{\AA}^2</math>)</b>	
Receptor	84.1
T4L	97
Risperidone	75.8
Lipids, water, other	86.8
<b>Model Statistics</b>	
RMSD-bonds ( $\text{\AA}$ )	0.004
RMSD-angles ( $^\circ$ )	0.56
Ramachandran favored (%) <sup>#</sup>	97.36
Ramachandran allowed (%) <sup>#</sup>	2.64
Ramachandran outliers (%) <sup>#</sup>	0.00
Rotamer outliers (%) <sup>#</sup>	0.67
Clashscore <sup>#</sup>	3.99

Highest resolution shell is shown in parentheses.

\* $R_{\text{merge}} = \sum hkl |I(hkl) - \langle I(hkl) \rangle| / \sum hkl I(hkl)$ , where  $\langle I(hkl) \rangle$  is the mean of the symmetry equivalent reflections of  $I(hkl)$ .

<sup>#</sup>As defined in MolProbity.

Extended Data Table 3 | Affinity of risperidone and nemonapride for ligand-binding-pocket mutants of the D<sub>2</sub> dopamine receptor

Receptor	Risperidone		Nemonapride	
	K <sub>i</sub> , nM (pK <sub>i</sub> ± SEM)	ΔpK <sub>i</sub> (mutant-WT)	K <sub>d</sub> , nM (pK <sub>d</sub> ± SEM)	ΔpK <sub>d</sub> (mutant-WT)
DRD2 wild-type	4.50 (8.41 ± 0.07)	--	0.21 (9.69 ± 0.06)	--
DRD2 W100 <sup>EL1</sup> A	8.14 (8.19 ± 0.13)	-0.21	1.97 (8.71 ± 0.02)	-0.98
DRD2 F110 <sup>3.28</sup> A	36.89 (7.48 ± 0.09)	-0.93	0.17 (9.77 ± 0.03)	0.08
DRD2 D114 <sup>3.32</sup> A	>10000	--	8.10 (8.09 ± 0.04)	-1.60
DRD2 V115 <sup>3.33</sup> A	3.07 (8.52 ± 0.04)	0.11	0.84 (9.08 ± 0.02)	-0.61
DRD2 C118 <sup>3.36</sup> A	4.84 (8.32 ± 0.01)	-0.09	0.40 (9.40 ± 0.02)	-0.29
DRD2 T119 <sup>3.37</sup> A	177.19 (6.83 ± 0.12)	-1.58	0.43 (9.38 ± 0.06)	-0.31
DRD2 I122 <sup>3.40</sup> A	13.87 (7.97 ± 0.13)	-0.44	0.30 (9.52 ± 0.01)	-0.17
DRD2 S197 <sup>5.46</sup> A	1.22 (8.92 ± 0.03)	0.51	0.43 (9.37 ± 0.01)	-0.32
DRD2 F198 <sup>5.47</sup> A	41.95 (7.38 ± 0.03)	-1.02	0.76 (9.12 ± 0.02)	-0.57
DRD2 F382 <sup>6.44</sup> A	57.70 (7.25 ± 0.05)	-1.16	0.30 (9.53 ± 0.05)	-0.16
DRD2 W386 <sup>6.48</sup> A	>10000	--	4.02 (8.40 ± 0.04)	-1.29
DRD2 F389 <sup>6.51</sup> A	2992 (5.65 ± 0.17)	-2.76	4.70 (8.35 ± 0.08)	-1.34
DRD2 F390 <sup>6.52</sup> A	31.20 (7.61 ± 0.15)	-0.80	1.30 (8.89 ± 0.03)	-0.80
DRD2 Y408 <sup>7.35</sup> A	13.63 (7.95 ± 0.13)	-0.46	0.18 (9.76 ± 0.02)	0.07
DRD2 T412 <sup>7.39</sup> A	102.68 (7.02 ± 0.08)	-1.77	4.92 (8.33 ± 0.10)	-1.36
DRD2 Y416 <sup>7.43</sup> A	2772 (5.61 ± 0.15)	-2.80	0.88 (9.06 ± 0.01)	-0.63

Data represent mean K<sub>i</sub> (pK<sub>i</sub> ± s.e.m.) for competition-binding experiments and K<sub>d</sub> (pK<sub>d</sub> ± s.e.m.) for homologous competition-binding experiments using [<sup>3</sup>H]-nemonapride (0.1–0.5 nM) as radioligand. All data are mean ± s.e.m. of three independent assays (n = 3 independent experiments).



Extended Data Table 4 | Compound dissociation and association rates on wild-type and mutant DRD2

Compound	Receptor	Residence Time, min ( $k_{off} \pm \text{SEM}$ ) min <sup>-1</sup>	$k_{on} \pm \text{SEM}$ , M <sup>-1</sup> min <sup>-1</sup>	$K_d$ , nM (pK <sub>d</sub> $\pm$ SEM)
Aripiprazole	DRD2 wild-type	154 (0.0065 $\pm$ 0.0004)	7.68 $\times 10^5$ $\pm$ 4.94 $\times 10^5$	9.43 (8.03 $\pm$ 0.07)
	DRD2 W100 <sup>E1</sup> A	15 (0.067 $\pm$ 0.015) <sup>*p&lt;0.006</sup>	2.48 $\times 10^5$ $\pm$ 6.5 $\times 10^4$	273 (6.56 $\pm$ 0.02)
	DRD2 W100 <sup>E1</sup> L	14 (0.071 $\pm$ 0.0007) <sup>*p&lt;0.006</sup>	1.89 $\times 10^5$ $\pm$ 3.7 $\times 10^4$	387 (6.42 $\pm$ 0.08)
	DRD2 W100 <sup>E1</sup> F	26 (0.038 $\pm$ 0.007) <sup>*p&lt;0.008</sup>	6.32 $\times 10^5$ $\pm$ 8.5 $\times 10^4$	62.8 (7.22 $\pm$ 0.14)
	DRD2 L94 <sup>E2</sup> E4A	59 (0.017 $\pm$ 0.002) <sup>n.s.</sup>	1.23 $\times 10^6$ $\pm$ 1.06 $\times 10^6$	49.8 (7.56 $\pm$ 0.52)
	DRD2 I184 <sup>E1</sup> E2A	100 (0.010 $\pm$ 0.001) <sup>n.s.</sup>	6.65 $\times 10^5$ $\pm$ 5.1 $\times 10^4$	15.5 (7.81 $\pm$ 0.03)
	DRD2 L94 <sup>E2</sup> E4A/I184 <sup>E1</sup> E2A	3 (0.32 $\pm$ 0.06) <sup>*p&lt;0.005</sup>	2.93 $\times 10^5$ $\pm$ 2.58 $\times 10^5$	413 (6.64 $\pm$ 0.52)
N-Methylspiperone	DRD2 wild-type	250 (0.004 $\pm$ 0.0003)	2.34 $\times 10^8$ $\pm$ 6 $\times 10^7$	0.018 (10.75 $\pm$ 0.08)
	DRD2 W100 <sup>E1</sup> A	21 (0.048 $\pm$ 0.0079) <sup>*p&lt;0.0073</sup>	1.65 $\times 10^8$ $\pm$ 6 $\times 10^7$	0.31 (9.51 $\pm$ 0.08)
	DRD2 W100 <sup>E1</sup> L	20 (0.050 $\pm$ 0.0064) <sup>*p&lt;0.0072</sup>	1.72 $\times 10^8$ $\pm$ 4 $\times 10^7$	0.29 (9.53 $\pm$ 0.03)
	DRD2 W100 <sup>E1</sup> F	38 (0.026 $\pm$ 0.00003) <sup>*p&lt;0.0083</sup>	2.08 $\times 10^8$ $\pm$ 5 $\times 10^7$	0.13 (9.89 $\pm$ 0.10)
	DRD2 L94 <sup>E2</sup> E4A	77 (0.013 $\pm$ 0.0047) <sup>n.s.</sup>	2.08 $\times 10^8$ $\pm$ 4 $\times 10^7$	0.062 (10.21 $\pm$ 0.08)
	DRD2 I184 <sup>E1</sup> E2A	128 (0.0078 $\pm$ 0.00004) <sup>n.s.</sup>	1.70 $\times 10^8$ $\pm$ 3 $\times 10^7$	0.048 (10.33 $\pm$ 0.08)
	DRD2 L94 <sup>E2</sup> E4A/I184 <sup>E1</sup> E2A	6 (0.170 $\pm$ 0.063) <sup>*p&lt;0.0064</sup>	1.62 $\times 10^8$ $\pm$ 1 $\times 10^7$	1.02 (9.01 $\pm$ 0.14)
Nemonapride	DRD2 wild-type	167 (0.006 $\pm$ 0.0002)	2.0 $\times 10^8$ $\pm$ 5 $\times 10^7$	0.031 (10.52 $\pm$ 0.09)
	DRD2 W100 <sup>E1</sup> A	43 (0.023 $\pm$ 0.001) <sup>*p&lt;0.002</sup>	1.17 $\times 10^8$ $\pm$ 2 $\times 10^7$	0.19 (9.75 $\pm$ 0.14)
	DRD2 W100 <sup>E1</sup> L	45 (0.022 $\pm$ 0.0018) <sup>*p&lt;0.003</sup>	1.07 $\times 10^8$ $\pm$ 3 $\times 10^7$	0.20 (9.70 $\pm$ 0.02)
	DRD2 W100 <sup>E1</sup> F	40 (0.025 $\pm$ 0.0019) <sup>*p&lt;0.002</sup>	2.03 $\times 10^8$ $\pm$ 6 $\times 10^7$	0.13 (9.90 $\pm$ 0.10)
	DRD2 L94 <sup>E2</sup> E4A	26 (0.039 $\pm$ 0.0033) <sup>*p&lt;0.0015</sup>	2.97 $\times 10^8$ $\pm$ 5 $\times 10^7$	0.13 (9.88 $\pm$ 0.03)
	DRD2 I184 <sup>E1</sup> E2A	149 (0.0067 $\pm$ 0.0004) <sup>n.s.</sup>	9.60 $\times 10^7$ $\pm$ 7 $\times 10^6$	0.07 (10.16 $\pm$ 0.06)
	DRD2 L94 <sup>E2</sup> E4A/I184 <sup>E1</sup> E2A	5 (0.20 $\pm$ 0.0048) <sup>*p&lt;0.0009</sup>	2.89 $\times 10^8$ $\pm$ 1 $\times 10^8$	0.82 (9.12 $\pm$ 0.19)

Data were acquired by association and dissociation kinetic experiments conducted in parallel at room temperature using [<sup>3</sup>H]-N-methylspiperone (0.8–1.0 nM) for aripiprazole and N-methylspiperone or [<sup>3</sup>H]-nemonapride (0.8–1.0 nM) for nemonapride. Estimates of  $k_{off}$ ,  $k_{on}$ , and  $K_d$  were obtained from four independent experiments. Residence time was calculated as  $1/k_{off}$ . All data are mean  $\pm$  s.e.m. of four independent assays ( $n = 4$  independent experiments). \*statistically significant differences between wild-type and mutant receptors; n.s., not significant;  $P$  values are indicated, unpaired two-tailed Student's  $t$ -test.

## CORRIGENDUM

doi:10.1038/nature25162

### Corrigendum: Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance

Sydney M. Shaffer, Margaret C. Dunagin, Stefan R. Torborg, Eduardo A. Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A. Brafford, Min Xiao, Elliott Eggan, Ioannis N. Anastopoulos, Cesar A. Vargas-Garcia, Abhyudai Singh, Katherine L. Nathanson, Meenhard Herlyn & Arjun Raj

*Nature* **546**, 431–435 (2017); doi:10.1038/nature22794

In Extended Data Fig. 4a of this Letter, the ‘JUN’ and ‘AXL’ labels on the histograms were inadvertently switched. The unmodified plot files and R code used to generate this figure are available on the Dropbox repository available with this Letter (link to the plot file: [https://www.dropbox.com/s/fl4ungg63rz4mnw/WM9\\_noDrug\\_20150810\\_geneHistogramRugs.pdf?dl=0](https://www.dropbox.com/s/fl4ungg63rz4mnw/WM9_noDrug_20150810_geneHistogramRugs.pdf?dl=0); link to R code: [https://www.dropbox.com/sh/g9c84n2torx7nuk/AABrhqLdqD8jGw7vd1duKXZ6a/paper/plotScripts/RNAFISH?dl=0&preview=RNAFISH\\_analysis.R](https://www.dropbox.com/sh/g9c84n2torx7nuk/AABrhqLdqD8jGw7vd1duKXZ6a/paper/plotScripts/RNAFISH?dl=0&preview=RNAFISH_analysis.R)). Extended Data Fig. 4a of the original Letter has been corrected, and the original panel is shown as Supplementary Information to this Corrigendum for transparency. In addition, in Extended Data Fig. 9a, the underlying image for ‘SK-MEL-28 NGFR sort biological replicate 2’ was duplicated and also shown for ‘WM989-A6 NGFR sort biological replicate 2’. The overlying colony counts for this dataset were correct. The Dropbox repository available with this Letter contains the correct images for both samples (link for SK-MEL-28 NGFR sort biological replicate 2: [https://www.dropbox.com/s/qek7o4l20tob33m/polygonImg\\_20160602\\_SKMEL\\_replicate.eps?dl=0](https://www.dropbox.com/s/qek7o4l20tob33m/polygonImg_20160602_SKMEL_replicate.eps?dl=0); link for WM989-A6 NGFR sort biological replicate 2: [https://www.dropbox.com/s/c0bajc19qfswghz/polygonImg\\_20160517.eps?dl=0](https://www.dropbox.com/s/c0bajc19qfswghz/polygonImg_20160517.eps?dl=0)). Extended Data Fig. 9a of the original Letter has been corrected, and the original panel is shown as Supplementary Information to this Corrigendum for transparency. We thank S. Y. Lee for identifying the duplicated image and bringing it to our attention. The conclusions of this Letter are unaffected by these errors.

**Supplementary Information** is available in the online version of this Corrigendum.

## CORRIGENDUM

doi:10.1038/nature25993

### Corrigendum: Landscape of X chromosome inactivation across human tissues

Taru Tukiainen, Alexandra-Chloé Villani, Angela Yen,  
Manuel A. Rivas, Jamie L. Marshall, Rahul Satija,  
Matt Aguirre, Laura Gauthier, Mark Fleharty, Andrew Kirby,  
Beryl B. Cummings, Stéphane E. Castel, Konrad J. Karczewski,  
François Aguet, Andrea Byrnes, GTEx Consortium,  
Tuuli Lappalainen, Aviv Regev, Kristin G. Ardlie, Nir Hacohen &  
Daniel G. MacArthur

*Nature* **550**, 244–248 (2017); doi:10.1038/nature24265

In this Letter, the Source Data associated with Fig. 2a and d were incorrect. This was due to an error during manuscript preparation, when transformed data instead of the raw values plotted in the figure were included in the Source Data file. The figure panels are correct and remain unchanged, and these errors do not affect the results or conclusions of the Letter. We apologize for any confusion this may have caused. The original incorrect Source Data for Fig. 2 are provided as Supplementary Information to this Corrigendum, for transparency. The original Letter has been corrected online.

**Supplementary Information** is available in the online version of this Corrigendum.

## RETRACTION

doi:10.1038/nature25779

### **Retraction: Asia's glaciers are a regionally important buffer against drought**

Hamish D. Pritchard

*Nature* **545**, 169–174 (2017); doi:10.1038/nature22062

In this Article, I estimated net glacial melt volumes on the river-basin scale from long-term precipitation and temperature records (1951–2007), taking into account the various mass contributions from avalanching, sublimation, snow drifting and so on. To this component (the seasonally delayed turnover of water in the glacial system) I added an estimate of the contribution due to sustained glacial mass losses, based on sparse observations of multi-decadal change. I then accounted for meltwater losses through evaporation, and compared this to net precipitation, distributed across river basins and across the catchments of a large number of dams. I estimated the second meltwater component (the additional contribution from glacier losses) as  $-0.35$  to  $-0.40$  metres water-equivalent per decade based on a global compilation of long-term mass-balance observations (from table 2 in ref. 32 of the Article). In this table, losses are described as “decadal averages (millimetres water equivalent)” but the units are actually intended to be decadal averaged annual values. Hence, the loss components of total meltwater that I used in my calculations are too small and the summed meltwater volumes reported here should be larger. Asia's glaciers are thus regionally a more important buffer against drought than I first stated, strengthening some of the conclusions of this study but also altering others. I am therefore retracting this Article. I thank L. Zhao and J. Moore for bringing the error to my attention.



# CAREERS

**INDUSTRY** Women in tech roles are feeling increasingly isolated **p.276**

**ACTIVISM** ‘March for Science’ organizers hope to repeat last year’s success **p.276**

**CAREER PATHS** Science PhDs remain greatly valued in the job market **p.277**

ILLUSTRATION ADAPTED FROM GETTY



## COLUMN

# Boost your market value

To get grants and jobs, know where your skills will be valued and how to promote them.

BY PETER FISKE

Most academic researchers might not be familiar with marketing, but my own long stint in the private sector and my recent return to academia have taught me that a market strategy can be a crucial step in winning grants, building a scientific reputation and advancing your career.

When I left academia in 2000 to launch my first company, I thought that marketing and sales were the same thing. To my mind, they both involved showing potential customers my company’s amazing technology and believing that they would want to buy it. But after several months, I hadn’t closed on a single sale.

Why? I hadn’t done a market analysis to learn whether my product would actually meet my customers’ needs. And I had not developed a market strategy to attract their attention.

A business-adviser colleague explained that sales involves presenting your product or service to prospective customers and addressing the decisions and steps that they must take before they buy. Marketing, by contrast, is a two-part process: figuring out what that product or service needs to be (in other words, carrying out a market analysis), and then working out how best to promote and present it (coming up with a market strategy).

### ADJUSTING THE FOCUS

I soon learnt that even the best technology does not ‘sell’ itself. I was certain that our high-precision optical tools were superior to those of our competitors, but I did not understand or appreciate what might keep potential customers from switching to our product.

So I interviewed customers to determine exactly what they lacked and to work out how we could provide it to them. Our sales took

off as soon as I had learnt that my clients were desperate for two things: rapid order fulfilment and test-certified optical components.

How does all this apply to scientists? Some might think that market analyses and market strategies have nothing to do with their work. Take grants, for example. Requests for grant proposals (RFPs) already specify the funder’s requirements and the boundaries of the research problem to be funded. It would seem that no market analysis is required.

But if you’re seeking funding from any source — whether a government agency, a foundation or a business — bear in mind that long before a funder issues an RFP, it assesses key areas of research need. Funders might organize invitation-only workshops to gather feedback from research or industry leaders. Often, a funding agency will have published a strategic plan or technical road map that identifies the priority areas for research. The US ►

## INDUSTRY

## Women alone in tech

US corporate-training programmes aimed at retaining female researchers in technology might be focusing on the wrong targets. A report published in February examines the results of in-depth interviews with 23 women in information-technology jobs across industry, including some at manufacturers, software-development firms and an insurance company (H. Annabi and S. Lebovitz *Inf. Syst. J.* <http://dx.doi.org/10.1111/isj.12182>; 2018). The authors sought to identify challenges faced by female researchers in this field. Employers often invest in female-centred mentoring and professional development, but the study participants said that they still feel forced out by their work environment. Fifteen respondents reported feeling isolated and excluded at work, and 13 said that a male-dominated workplace causes feelings of alienation. “There’s a mismatch with these investments in training and the barriers that women actually face,” says lead author Hala Annabi, an information-systems scholar at the University of Washington in Seattle. A Pew Research Center report (see [go.nature.com/2esrhz5](http://go.nature.com/2esrhz5)) found that the proportion of women in computer-related fields in the United States has dropped from 32% in 1990 to 25% today.

## SCIENCE ACTIVISM

## March for advocacy

The second March for Science is scheduled for 14 April in Washington DC ([marchforscience.com](http://marchforscience.com)). Organizers hope to recapture the energy and enthusiasm of last year’s event, when more than 1 million researchers and others — in 600 cities around the world — marched in support of evidence-based policy and the application of science for the greater good. Organizers worldwide expect events with fewer marchers, placards and chants but more advocacy-related activities. Berlin is planning a ‘local hero’ programme in which scientists will give public talks at cafes and other venues. March-related activities in Portland, Oregon, will include speeches by local politicians and a science expo with at least 30 presenters. The election and inauguration of Donald Trump as US president helped to spur marchers last year. But Caroline Weinberg, an organizer of the march, says that science activism shouldn’t depend on controversial events to draw interest and participation. “We can’t allow our advocacy to be tethered to those moments,” she says.

## PLAY TO YOUR STRENGTHS

## Marketing tips for a job search

- Understand yourself. List your key technical skills, experience, perspective and approach to problem-solving. What problems do you solve best, and in which situations or environments do you produce your best work? When have you been your happiest at work and what were you doing? Knowing this will help you to identify the types of employer for whom you can add the greatest value.
- Conduct a market analysis for the jobs and fields that interest you. Seek out people who received their PhD in the same field as yours, or in one that’s similar, but who have gone in different professional directions. Ask them where scientists with your background and strengths have been successful. Identify industries in which your skills and experience are relevant and valued, and investigate organizations whose mission aligns with your work. Gain a ‘market perspective’ on an industry

by joining a professional organization or taking a short course or workshop to understand how your scientific background might align with that interest.

- Expand your network. Reach out each week to people in positions that interest you, and meet them in person, if possible, to learn more about what they do. Follow up with them periodically to let them know your professional trajectory. Not only will you gain insights into positions or roles that interest you, but you might get help from these contacts in your job search.
- Focus on opportunities. Identify those organizations that you feel are the best fit for your skills, interests and values. Conduct informational interviews with key managers — who may be expanding their teams in the future — to get a feel for the work environment. Find out the managers’ goals and needs and see how your skills and background could help. **P.F.**

► Department of Energy’s Advanced Manufacturing Office, for example, has published a five-year plan (see [go.nature.com/2elyc71](http://go.nature.com/2elyc71)). You should review such documents, as well as past RFPs from the agency concerned, and aim to learn from colleagues or associates what took place at earlier planning workshops.

If you don’t personally know former programme managers at an agency, you can search for them on LinkedIn, and find out which research ideas overall have proved most successful at that agency. And you can contact the funder itself and speak to a grant administrator or programme manager to learn whether your specific research idea pertains to the funder’s strategic interest (see *Nature* **482**, 429–431; 2012). (Grant-writers should first study a funder’s website and grant materials to learn the funder’s priorities, and glean background information and context.)

## STREAMLINE YOUR SEARCH

Many early-career scientists fail to conduct a market analysis or develop a market strategy for their job search. They wait for a job advertisement to appear and then submit a CV — a compendium of every element of their research career so far — and hope that their background and research experience will merit further review.

Instead, before applying for specific jobs, you should deploy the market analysis-and-strategy template outlined above (see ‘Play to your strengths’). Sound out people who are already working in a field or for organizations that interest you (see *Nature* **538**,

417–418; 2016). Ideally, aim to connect with scientists whose backgrounds are similar to yours — perhaps they earned a PhD in the same field or from the same institution — and who have enough experience to directly advise you on where your skills and interests fit, and how best to present yourself.

As part of your market strategy, you should also craft and maintain a profes-

sional online persona. Use a platform such as LinkedIn or ResearchGate to create a detailed profile emphasizing key skills and experience, and to link up with others in relevant organizations or fields of research.

Use online technical forums to ask about skills and experience needed in an industry or for a specific position (part of your market analysis), and answer technical questions posed by others. Taking part in such dialogues can make recruiters notice you and seek you out regarding prospective openings.

These marketing activities are time-consuming. But they offer crucial insight into where a discipline or a field of technology is heading, and into the skills, knowledge and experience that you’ll find most valuable. ■

**Peter Fiske** is director of the Water-Energy Resilience Research Institute at the Lawrence Berkeley National Laboratory in Berkeley, California.

## PROFESSIONAL OUTCOMES

# PhD career paths hold promise

*Most people studying for science doctorates land a job that they enjoy after graduating.*

BY CHRIS WOOLSTON

As universities around the world award science PhDs at an ever-increasing rate, some doctoral students might wonder whether the degree is still worth all the time, effort and sacrifice.

But two recent projects tracking the journeys of PhD holders in the United Kingdom and Canada offer reason for optimism: graduates in the sciences and other fields are highly employable, even if they don't always end up where they expected. "There's a lot of pessimism about an oversupply of PhDs," says Sally Hancock, an education researcher at the University of York, UK, who led the study in her nation — one of only a few of its kind worldwide. "These data can help demystify what happens next."

Using information collected by the UK Higher Education Statistics Agency, Hancock analysed the job outcomes of more than 4,700 people throughout the United Kingdom who graduated with a PhD in either the 2008–09 or 2010–11 academic years. All respondents were surveyed 3.5 years after graduation.

Hancock's analysis, funded by the UK Society for Research into Higher Education and yet to be published, suggests that around 2% of graduates across all fields were unemployed, and nearly 80% had full-time jobs. Close to 10% worked part-time. The rest were mostly pursuing further studies or doing volunteer work.

Nearly 30% of those with full- or part-time jobs ended up in academia. Of those, about 70% worked as teaching professionals and 30% were university researchers. Around 20% worked in industry, often as researchers or managers. Another 20% held medical jobs, including as practitioners and medical scientists.

PhD holders, at least in the United Kingdom, are hardly on the poverty line, says Janet Metcalfe, head of Vitae, a non-profit science-career advocacy organization in Cambridge, UK. "It's been like that since the 1970s," Metcalfe says. "They've always been highly employable. They've always had a premium over those who hold master's and undergraduate degrees."

Previous Vitae surveys, Metcalfe notes, have found that roughly 80% of postdocs want to remain in academia — many more than actually do so (see *Nature* 550, 549–552; 2017). "There's a complete mismatch between career aspirations and the potential for getting academic positions," she says.

Although many people with PhDs end up changing course from their original career plan, that hasn't drastically eroded career satisfaction:

more than 95% of respondents across all sectors in Hancock's analysis said that they were at least somewhat satisfied with their careers, including 48% who said they were very satisfied. "Satisfaction doesn't vary much by sector," Hancock says. "Even if it's not what they expected, the outcomes are OK."

Hancock's analysis revealed some disparities in salaries. Those reported for graduates in academia (a yearly median of £37,000, or US\$51,000) were higher than those in industry (£36,000 for men; £34,000 for women).

Women in the biological sciences reported earning a median of £35,000 per year compared with £36,000 for men. The gender gap was slightly larger in the physical sciences and engineering, where women reported a median salary of £34,000 and men £36,000. The biggest gap was in the biomedical sciences, where women reported an annual salary of £36,000, whereas men earned £45,000. "There are persistent and stubborn gender differences," Hancock says, but she adds that the data offer no clues about the root cause of the pay disparities.

Metcalfe says that the data do not make it clear whether UK female scientists are getting short-changed. She notes that the survey used salary ranges, not exact salaries, and that the relatively small number of people surveyed in the biomedical sciences — fewer than 600 — makes the figures sensitive to outliers.

In Canada, the 10,000 PhDs Project at the University of Toronto (UT) (see [go.nature.com/2ccdzyj](http://go.nature.com/2ccdzyj)), led by biochemist Reinhart Reithmeier, also found encouraging results. The project tracked outcomes for all PhD holders who received a doctoral degree from UT between 2002 and 2015. Through online searches, project researchers verified job titles for 9,583 PhD holders, or 88% of all graduates. The study has no data for the remaining 12%, but Reithmeier notes that in the 2016 census, the unemployment rate for all PhD graduates in Canada was 5.1%.

Science doctoral degrees led to a wide array of positions in Canada. About 23% of respondents have tenure or tenure-track positions, and

just over half work in any type of academic position, including as administrators. Nearly 30% are in industry, and others work for the federal or provincial governments, charities or entrepreneurial businesses.

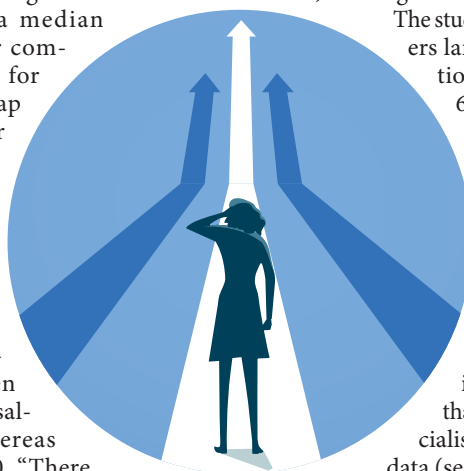
The unusually large percentage of graduates in academia might be a local phenomenon, says Joshua Barker, dean of UT's School of Graduate Studies. "We know that a lot of our graduates like to stay in the region," he says. The report found that the city's two largest universities — UT and York University — employed nearly 1,200 UT graduates between them.

The study shows that PhD holders landed a variety of positions in industry. Nearly 60% of life-sciences graduates now working in the private sector ended up in biotechnology or pharmaceutical jobs. But 13% of all physical-sciences PhDs in the private sector work in banking, finance or investments — sectors that increasingly need specialists who can manage big data (see *Nature* 548, 613–614; 2017). "These niches probably didn't exist 15 years ago," says Reithmeier.

The UT findings were largely consistent with a survey from the University of British Columbia in Vancouver done in 2016, which tracked graduates who had earned PhDs from 2005 to 2013 (see [go.nature.com/2tdcgh9](http://go.nature.com/2tdcgh9)). Just over half of those graduates had positions in academia; of those, nearly 15% had postdoctoral fellowships. More than 91% of survey respondents said that they felt as if they were on the right career path, but some reported that they felt overqualified, unable to find work that was relevant to their doctoral degree. "I don't want to ignore those who are struggling and unhappy," says Susan Porter, the university's dean and vice-provost of graduate and postdoctoral studies. "Some feel that they were fed a line."

The main takeaway, Metcalfe says, is that PhD recipients should feel confident in their career potential, especially if they are willing to look beyond universities. "All of our language in academia encourages researchers to be academics," she says. "The challenge is getting over this psychological barrier to help researchers look more widely in terms of employment. There are some great jobs out there." ■

ILLUSTRATION ADAPTED FROM GETTY





# WRITING FOR THE END OF THE WORLD

## A mammoth task.

BY KARLO YEAGER RODRÍGUEZ

What apocalypse am I creating today?

I feed two blank sheets of onionskin, carbon paper sandwiched between, through my typewriter's platen. I stop a moment, remembering this joke. *This guy gets on a bus*, it starts, but I don't write that.

Instead, I begin my 302nd story about the end of the world. I stretch my fingers over the keys, looking out the window. When I first got this cabin, I chose it for its view of the pond, cradled by the mountains. Today, I'm choosing it for its isolation from the world.

Off the grid. Solar panels on the roof; overflow to shielded battery back-ups underground. Enough to run my fridge. Wood-burning stove, in case I'm up here during the winter. Battery-powered well pump, with a UV sanitation system for if the pond gets contaminated. Hell, I even have an e-reader with a downloaded version of Wikipedia to make sure I can still do research.

I ache to turn on my radios — Internet app and ham, if the other goes offline — and see what's happening out in the world, but no.

I have a deadline.

So, this guy, he gets on a bus. Across from where he sits, is another guy gripping a battered briefcase on his lap, a thin sheen of sweat on his face. Every time the bus stops, nervous briefcase man cracks open his briefcase and throws fistfuls of crumpled paper out the window.

The first time I ended the world was at a workshop down near DC. They wanted to contract me as a — get this — futurist. I'd only ever written some stories with rockets and ray-guns, but the pay was good.

Q: What's the difference between 'writer' and 'futurist'?

A: The amount of dollars in front of the decimal.

If I accepted, I was to get myself to Union Station, where they would transport me the rest of the way to an undisclosed location. When I approached the young man holding a card with my name on it, he directed me towards an olive-drab bus. Another author flumped down into the seat across from mine. Instead of putting his briefcase under his seat, he kept it in his lap. He caught my

glance, and patted its leather.

"Gift from my daughter," he said. "To celebrate my first sale." I nodded and smiled — it was the polite thing to do — but didn't think it important at the time.

We pulled into a corporate office park somewhere in Alexandria, and filed off into the building. Bad coffee, good doughnuts. I balanced one atop the other before being



led into a large classroom with long tables.

Apart from a huge mirror where a chalkboard would be, the room was as bland and forgettable as our instructor. As he explained different scenarios, I stared at the mirror. I was sure the spooks and top brass were watching us from behind it.

"Be creative when writing about apocalypses," he said.

*Apocalypses.*

Plural.

I always assumed there would be only one.

"Imagine the danger is *imminent*." He paused to look at all of us. "Because it is. Think it all through.

"Worst-case, best-case and all scenarios in between. Everyone is affected, whether it's the leader of a country or of the local PTA. We want you to think of all types of stories."

"Are *we*," briefcase author said, "actually in danger? I can't help but notice *we* were the ones who dropped the bomb on another country."

Others shrank away from him, and silence filled the space.

Was it lunchtime? Dinner?

We had all been under the flicker of fluorescents so long, nobody knew any more. In a workshop run by spooks, were they studying whom among us could do what they needed?

Back to the joke.

So, our guy notices nervous briefcase man throwing wads of paper from his briefcase out the window at the first stop. Then the

second, and on the third, he decides to find out why he's doing it.

"Excuse me, sir," he says. "I can't help but notice every time we stop, you open your briefcase, and throw papers out the window."

"That's right," nervous briefcase man says after a minute.

"Why?"

At the question, nervous briefcase man's face lights up with a sly smile, and he beckons our guy a bit closer.

"It's to keep the elephants away," he confides.

They taught us how the disruption of human systems can cause the end of the world. As I worked, the other writers, clicking their pens, or drumming their fingers, imagined their loved ones stricken down in the first wave, or left to lingering deaths. Each wrestling with their emotions.

Long enough to pin them down, observe and describe them.

The writer with the briefcase, sat alone at his desk, dabbing at his upper lip while taking notes. Back on the bus, he muttered in the dark, hugging his briefcase to his chest.

He never made it off the bus.

Our instructor intervened, and prevented the police from taking our statements. Whatever he told them made the Capitol Police retreat to their patrol cars. They debriefed us on the incident.

He'd had a bad ticker, if you can believe it.

"But, sir," our guy on the bus says, "there aren't any elephants for thousands of miles!"

The bus is coming up to the next stop, and nervous briefcase guy cracks the case open, crumpling papers into a fist.

"See how well it works?"

I start typing, eyes on the distant mountains.

Hope is important in these stories. No matter how grim things can get, I always try to leave room for it. People like stories like that.

When I'm done, I peel the carbon copy off, stuff it in an envelope to be mailed out to a magazine. The original, I put in my safe, hidden in the crawlspace.

Will this story stop the apocalypse this time?

I doubt it, but I've got to try something. ■

**Karlo Yeager Rodríguez** is from the enchanted island of Puerto Rico, but moved to Baltimore some years back. He lives happily with his partner and one very odd dog.

ILLUSTRATION BY JACEY



# natureOUTLOOK

## THE FUTURE OF MEDICINE

8 March 2018 / Vol 555 / Issue No 7695



Cover art: Michele Marconi

### Editorial

Herb Brody,  
Michelle Grayson,  
Richard Hodson,  
Elizabeth Batty

### Art & Design

Mohamed Ashour,  
Andrea Duffy,  
Wesley Fernandes,  
Wojtek Urbanek

### Production

Mosud Ali, Ian Pope,  
Karl Smart

### Sponsorship

Janice Stevenson,  
Reya Silao,  
Anushree Roy

### Marketing

Nicole Jackson

### Project Manager

Rebecca Jones

### Art Director

Kelly Buckheit Krause

### Publisher

Richard Hughes

### Editorial director

Stephen Pincock

### Magazine Editor

Helen Pearson

### Editor-in-Chief

Philip Campbell

From the nineteenth-century benches of microbiologists Louis Pasteur and Robert Koch to the sequencing of the human genome in 2003, the past 200 years have seen medicine advance at an extraordinary pace. People are now enjoying longer and healthier lives than their ancestors. But as any medical researcher will attest, ambitions go much further.

The stories in this Outlook, chosen in consultation with editors from the Nature Research journals, represent some of the biggest opportunities we have to improve our future health. Our selection is not exhaustive, nor can we be certain that all research we report will come to fruition. But if only a fraction does, humanity can look forward to a healthier future.

To avoid antibiotic resistance undoing a century's worth of progress, researchers are racing to restock the antibacterial armoury (see page S5). Others are exploiting the data generated by ubiquitous computers and smartphones to better anticipate outbreaks of infectious disease (S2).

With the potential for gain so great, the prevention of illness is playing an ever-larger part in medicine (S20). Intervention to protect people from long-term disease could begin in the first moments after birth (S18). And although a decline in health in later life might seem normal, there is ongoing debate about where healthy ageing ends and disease begins (S15).

Work to exert greater control over rogue immune systems (S8), as well as to develop technological solutions to paralysis (S12), is showing initial promise. The advent of CRISPR–Cas genome editing has raised hopes for widespread use of gene therapy (S23); meanwhile, this technology is also aiding the search for new drugs (S10). As long as barriers to accessing the best treatments available can be negotiated away (S26), the future of medicine could be very bright indeed.

We are pleased to acknowledge the financial support of Merck in producing this Outlook. As always, *Nature* has sole responsibility for all editorial content.

**Richard Hodson**  
*Supplements editor*

## CONTENTS

### S2 BIG DATA

#### **Cloudy with a chance of flu**

Ways to improve outbreak forecasting

### S5 MICROBIOLOGY

#### **One step ahead**

Four advances in overcoming antibiotic resistance

### S8 IMMUNOLOGY

#### **Teaching tolerance**

Taming the immune system to help fight autoimmunity

### S10 PHARMACEUTICALS

#### **A CRISPR path to drug discovery**

A revolution in the search for new drugs

### S12 BIOENGINEERING

#### **The power of thought**

Restoring movement through neural prostheses

### S15 AGEING

#### **Lifting the burden of old age**

When does age-related muscle loss become a disease?

### S18 MICROBIOTA

#### **Baby thrivers**

How microbes met in early life might shape future health

### S20 PREVENTIVE MEDICINE

#### **Cleaning up our future health**

Limiting exposure to toxic chemicals

### S23 GENE EDITING

#### **The heart-disease vaccine**

A fresh approach to tackling a common condition

### S26 HEALTH ECONOMICS

#### **Cancer's cost conundrum**

Evaluating the price of oncology drugs

*Nature Outlooks* are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at [go.nature.com/e4dwz](http://go.nature.com/e4dwz)

### CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2018).

### VISIT THE OUTLOOK ONLINE

The *Nature Outlook The future of medicine* supplement can be found at [www.nature.com/collections/future-of-medicine-outlook](http://www.nature.com/collections/future-of-medicine-outlook). It features all newly commissioned content as well as a selection of relevant previously published material that is made freely available

for 6 months.

### SUBSCRIPTIONS AND CUSTOMER SERVICES

Site licences ([www.nature.com/libraries/site\\_licences](http://www.nature.com/libraries/site_licences)): Americas, [institutions@natureny.com](mailto:institutions@natureny.com); Asia-Pacific, <http://nature.asia/jp-contact>; Australia/New Zealand, [nature@macmillan.com.au](mailto:nature@macmillan.com.au); Europe/ROW, [institutions@nature.com](mailto:institutions@nature.com); India, [npgindia@nature.com](mailto:npgindia@nature.com). Personal subscriptions: UK/Europe/ROW, [subscriptions@nature.com](mailto:subscriptions@nature.com); USA/Canada/Latin America, [subscriptions@us.nature.com](mailto:subscriptions@us.nature.com); Japan, <http://nature.asia/jp-contact>; China, <http://nature.asia/china-subscribe>; Korea, [www.natureasia.com/ko-kr/](http://www.natureasia.com/ko-kr/) subscribe.

### CUSTOMER SERVICES

[Feedback@nature.com](mailto:Feedback@nature.com)

Copyright © 2018 Macmillan Publishers Ltd. All rights reserved.



world,” says Cécile Viboud, an epidemiologist at the US National Institutes of Health Fogarty International Center in Bethesda, Maryland. “It’s several orders of magnitude less than what we have in other fields.”

This year marks the centenary of the start of the Spanish flu pandemic, which involved a strain of flu virus known as H1N1 that killed up to 5% of the world’s population. The world is now much better prepared for such threats, as shown by the international reaction to the H1N1 pandemic of 2009, which was coordinated by a global network of laboratories that perform clinical testing. Yet the response was not swift enough to fully contain the pandemic, which claimed the lives of about 250,000 people in the first 12 months (F. S. Dawood *et al. Lancet Infect. Dis.* **12**, 687–695; 2012).

Lawrence Madoff, an infectious-disease specialist at the University of Massachusetts Medical School in Worcester, sees such delays as being an inherent constraint of conventional lab-based surveillance strategies. “They’re limited by their tendency to have rigid structures and count specific cases, and by a bureaucratic slowness that gets built into the system,” he says. For example, flu surveillance in the United States relies on a network called the Influenza-like Illness Surveillance Program, through which health-care providers across the country file weekly reports of probable cases on the basis of symptoms, and submit samples from patients to testing centres. The results are assessed centrally by the US Centers for Disease Control and Prevention (CDC). Consequently, even for a well-studied disease such as flu, it can take weeks to identify and respond to an outbreak. For diseases that are not monitored routinely, the delay can be catastrophic. For example, the response to the 2014–15 outbreak of Ebola in West Africa was described by an international panel of public-health specialists as an “egregious failure”, owing to the months-long delay before the World Health Organization (WHO) moved to contain what was already a full-blown emergency.

The good news is that the present era of widespread access to the Internet and digital health has created a rich reservoir of valuable data for researchers to dive into. “You could start to harness all this data that’s being generated on the web, gathered across different sources, to understand population health patterns,” says John Brownstein, a computational epidemiologist and chief innovation officer at Boston Children’s Hospital in Massachusetts. By harvesting and combining these streams of big data with conventional ways of monitoring infectious diseases, the public-health community could gain fresh powers to catch and curb emerging outbreaks before they rage out of control.

#### GOING VIRAL

Data scientists at Google were the first to make a major splash using data gathered online to track infectious diseases. The Google Flu

#### BIG DATA

# Cloudy with a chance of flu

*Internet search data, medical records and networks of on-the-ground experts could enable the accurate forecasting and faster control of disease outbreaks.*

BY MICHAEL EISENSTEIN

Even though you know it’s a sensible idea, you’re on the fence about whether it would be worth the bother to have this season’s influenza vaccine. But a quick glance at the flu forecast on your phone sets you straight: there’s a warning about a recent spike of cases

nearby, so you head to the clinic rather than risk a feverish week in bed. Epidemiologists eagerly anticipate such a future, in which they can track infectious diseases with the same confidence as meteorologists mapping the weather. But those making predictions of this type face a serious problem. “There is just not a lot of observational data in the disease



Trends algorithm, launched in November 2008, combed through hundreds of billions of users' queries on the popular search engine to look for small increases in flu-related terms such as symptoms or vaccine availability. Initial data suggested that Google Flu Trends could accurately map the incidence of flu with a lag of roughly one day. "It was a very exciting use of these data for the purpose of public health," says Brownstein. "It really did start a whole revolution and new field of work in query data."

Unfortunately, Google Flu Trends faltered when it mattered the most, completely missing the onset in April 2009 of the H1N1 pandemic. The algorithm also ran into trouble later on in the pandemic. It had been trained against seasonal fluctuations of flu, says Viboud, but people's behaviour changed in the wake of panic fuelled by media reports — and that threw off Google's data. "Before, only people who had flu were searching for flu symptoms," says Nicholas Generous, a biosurveillance researcher at Los Alamos National Laboratory in New Mexico. "All of a sudden, people that didn't have flu were searching and that ended up giving a false result." The project never recovered because "people at Google felt that it was not worth trying to improve the algorithm," says Viboud. The company stopped supporting Google Flu Trends in August 2015, although it continued to furnish academic and governmental organizations with relevant search data. "Google was a trailblazer, but monitoring diseases was not its primary purpose," says Viboud.

Nevertheless, its work with Internet usage data was inspirational for infectious-disease researchers. A subsequent study from a team led by Cecilia Marques-Toledo at the Federal University of Minas Gerais in Belo Horizonte, Brazil, used Twitter to get high-resolution data on the spread of dengue fever in the country. The researchers could quickly map new cases to specific cities and even predict where the disease might spread to next (C. A. Marques-Toledo *et al.* *PLoS Negl. Trop. Dis.* 11, e0005729; 2017). Similarly, Brownstein and his colleagues were able to use search data from Google and Twitter to project the spread of Zika virus in Latin America several weeks before formal outbreak declarations were made by public-health officials. Both Internet services are used widely, which makes them data-rich resources. But they are also proprietary systems for which access to data is controlled by a third party; for that reason, Generous and his colleagues have opted instead to make use of search data from Wikipedia, which is open source. "You can get the access logs, and how many people are viewing articles, which serves as a pretty good proxy for search interest," he says.

However, the problems that sank Google Flu Trends still exist. "Internet data is really great for diseases that are seasonal, where a lot of people get sick and there isn't a lot of media

hype," says Generous. "It probably wouldn't work for Ebola." He also notes that there are challenges in interpreting how people engage with the Internet on infectious diseases: they might be worried about their own symptoms, but could also have concerns about friends or family in high-risk areas, or simply be curious. Additionally, online activity differs for infectious conditions with a social stigma such as syphilis or AIDS, because people who are or might be affected are more likely to be concerned about privacy. Appropriate search-term selection is essential: Generous notes that initial attempts to track flu on Twitter were confounded by irrelevant tweets about 'Bieber fever' — a decidedly non-fatal condition affecting fans of Canadian pop star Justin Bieber.

Alternatively, researchers can go straight to the source — by using smartphone apps to ask people directly about their health. Brownstein's team has partnered with the Skoll Global Threats Fund to develop an app called Flu Near You, through which users can voluntarily report symptoms of infection and other information. "You get more detailed demographics about age and gender and vaccination status — things that you can't get from other sources," says Brownstein. Ten European Union member states are involved in a similar surveillance programme known as Influenzanet, which has generally maintained 30,000–40,000 active users for seven consecutive flu seasons. These voluntary reporting systems are particularly useful for diseases such as flu, for which many people do not bother going to the doctor — although it can be hard to persuade people to participate for no immediate benefit, says Brownstein. "But we still get a good signal from the people that are willing to be a part of this."

**"Internet data is really great for diseases that are seasonal, where a lot of people get sick."**

#### NETWORK NEWS

Internet activity and even self-reported data still leave a lot of room for interpretation. But front-line media reports can offer more trustworthy data points for signals of infectious diseases. One of the earliest forays into online epidemiology was ProMED-mail, established in 1994 as a mailing list for public-health experts to share reports of infectious diseases — including news stories, public-health announcements and clinical observations — from around the world. ProMED-mail blossomed rapidly into a widely used service that is managed by the International Society for Infectious Diseases in Brookline, Massachusetts. "More than 70,000 people now use it, and it's become a much more organized system of moderated reports," says Madoff, ProMED-mail's editor.

The service has also spawned a more

extensive effort known as HealthMap, an online atlas of infectious-disease reports built by Brownstein and his colleagues that pulls in data from ProMED-mail, reports from organizations such as the WHO and online news aggregated by Google and its Chinese counterpart Baidu. "All these news sites are out there," says Brownstein. "If you can just organize them, you can do an even better job of bringing down the time required to understand when a disease is unfolding." HealthMap extracts data automatically from these sources in real-time, giving it the advantage of speed in terms of catching a signal. But as with other attempts to computationally filter data from the Internet, researchers must be cautious of false positives such as mistaking news of malaria-related research for actual outbreaks of the disease. Accordingly, Madoff favours manual oversight for ProMED-mail. "Everything is hand-curated," he says.

When used properly, these Internet data streams can give the public-health community a head start in mobilizing a response to an outbreak. Madoff notes that ProMED-mail has pushed a number of emerging diseases into the public eye, compelling governments to take action. "We were first to report on MERS [Middle East respiratory syndrome] in Saudi Arabia in 2012," he says. "The Saudi Ministry of Health quickly responded and told us they knew about it and had a couple of other cases, and gave us more formal verification." A similar scenario played out for the outbreak of severe acute respiratory syndrome (SARS) in 2003, in which the Chinese government was initially reluctant to acknowledge the threat. "Once it was made public, they were ready to respond and became much more transparent," says Madoff. As a result, the international research community could begin to develop vaccines and treatments. And for known threats, HealthMap has outpaced conventional surveillance platforms in identifying recent infectious-disease events, including the outbreaks of both the H1N1 flu strain and Ebola. "We've shown these sources can bring down the time of detection by days or even weeks," says Brownstein.

#### TRUST, BUT VERIFY

All strategies for the indirect surveillance of disease still need to be clinically validated. This puts digital epidemiologists back under the constraints of conventional lab-based surveillance, which means weeks of delay — while patients seek medical care and samples are tested — before researchers can validate their signal. For less common diseases, this can be especially problematic. "The traditional surveillance data often is not there, or it's there but very patchy," says Viboud. Researchers are therefore looking to data from medicine's front line that are more reliable indicators of infectious-disease events.

Working directly with medical records is



Nicholas Generous and colleagues review data as part of efforts to forecast the spread of dengue fever.

one potential solution. In 2014, Viboud and her colleagues collaborated with medical-data company IMS Health (now part of Durham, North Carolina-based IQVIA), which provided de-identified medical claims filed across the United States. Analysis of the documents, which clinicians must submit to obtain reimbursement from health-insurance companies, produced a weekly view of flu transmission in US cities that was more detailed than the state-level data that are normally reported. “Medical claims are very solid because they’re based on actual visits to practitioners,” says Viboud. “To me, this is the most high-resolution data set out there.” However, such reports are also affected by delays, with many doctors filing claims weeks after seeing a patient, which makes the process better suited to post hoc epidemic analysis than to real-time surveillance.

Electronic health records could prove more useful as timely indicators of an outbreak, by helping to catch cases at the time of diagnosis. But using them poses privacy challenges — and in the United States, these data are under the control of private entities rather than a government agency, making it trickier to negotiate access. “They probably won’t make that data available to researchers — it will probably just be available to public-health officials,” says Generous. That would require the relevant local or national public-health agencies to act as intermediaries in processing and distributing health-record-derived insights to researchers, who can then use them in the modelling and analysis of epidemics.

This approach is limited to nations whose health-care systems are highly digitized, which is generally not the case for the low-income countries that have the highest burden of infectious disease. Madoff and his colleagues are trying to address this challenge through a programme called EpiCore — an army of

epidemiologists with Internet access that can be mobilized to confirm reports of infection directly. “We have over 2,000 volunteer epidemiologists now in around 140 countries who agree to be contacted in the event that there is an outbreak somewhere and to try to verify it,” says Madoff. “They can do so through an online platform that allows them to remain unidentified, so we can help people who are fearful of a government crackdown or something like that.”

For now, such diagnoses are being made the old-fashioned way, with health-care workers dispatching blood and other samples from people with symptoms to dedicated labs for testing. However, rapid strides in DNA-sequencing technology are making it feasible to achieve the accurate, on-site identification of pathogenic agents at minimal cost. Soon, it could be common for mobile diagnostic labs to acquire and upload genome data in the field. In 2016, for example, an international team of researchers took to the back roads of Brazil with a low-cost, portable sequencing system developed by Oxford Nanopore Technologies, based in the United Kingdom, which enabled them to analyse samples from across northeastern Brazil at the height of the Zika virus outbreak. Such an approach could tell public-health researchers exactly which strain of pathogen they’re grappling with, as well as help them to reconstruct chains of transmission — valuable information for containing and controlling infectious diseases. “We’ll get direct viral confirmation,” says Brownstein. “I’m not sure how long that’s going to take, but it will definitely replace what we’ve been doing up until now.”

#### PANDEMIC-PROOF

Another challenge will be to move beyond one-off demonstrations based on single data streams to a proven system that integrates

several sources of data — for example, coupling early warnings from ‘noisy’ social-media data with high-confidence signs of infection gleaned from hospital records — and that can be trained to pick up signals for several diseases at a time. “You might have a little bit of laboratory or clinical data that you can mix with Google Trends data or participatory surveillance,” says Viboud. “That’s where the field is going.” Veterinary data will also become an important piece of the puzzle, with the potential to give researchers warnings of emerging pathogens, and Madoff notes that ProMED-mail has included disease reports from livestock and wildlife since its inception. “You have to keep an eye on other species to know what might happen next in humans,” he says.

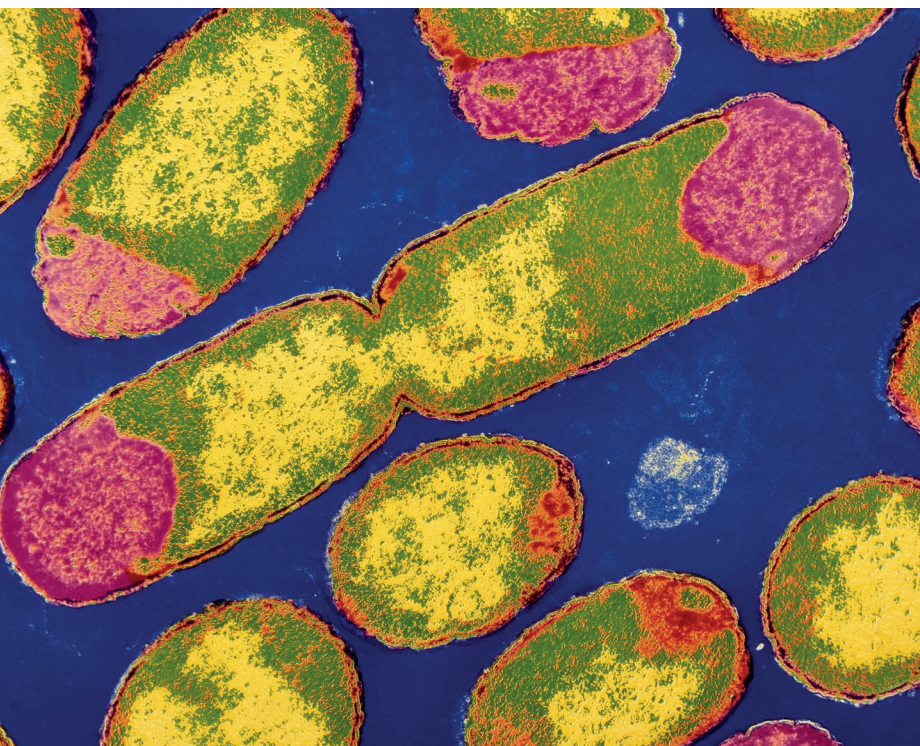
Tying these parts together will be difficult, not just because the various data sets quantify information at different scales of space and time, but also because nobody knows which combinations will improve public health. “If people actually do implement this operationally, what are the cost savings and life savings and health savings?” asks Generous. “Right now, it’s still a bit of throwing in everything but the kitchen sink and seeing what happens.” Without clear signs of their value to public health, such forays into digital epidemiology are likely to remain little more than intriguing experiments.

Yet the early evidence strongly suggests that for at least a handful of well-studied diseases, clever wrangling of data can buy the medical community extra days or weeks in which to act — time that could be used to quarantine unwell people or to mobilize clinicians or vaccine stocks. Public-health authorities are taking the idea seriously — since 2013, the CDC has run the ‘Predict the Influenza Season Challenge’ to stimulate research into outbreak forecasting. Viboud notes that the WHO has increased its focus on digital surveillance strategies following its heavily criticized response to Ebola. “The WHO hopes to get a network of modellers around the world that can help it for the next crisis,” she says.

Generous hopes that these efforts will ultimately transform into a resource for the public, enabling people to become informed consumers of epidemiological data and to take charge of their health in the same way that they might respond to information about traffic or the weather. The result could be a more sophisticated understanding of disease risk, guided by reality rather than media hype — although some education of users will be required. “When people first started to forecast weather, the idea that there was a ‘20% chance of rain’ must have been such a paradigm shift to understand, but we all get it now,” he says. “The question is, how does that happen for disease forecasting, and how does that become a routine, everyday thing?” ■

**Michael Eisenstein** is a freelance science writer based in Philadelphia, Pennsylvania.





The bacterium *Escherichia coli* is a common source of infection in the gut.

# ONE STEP AHEAD

Old drugs and new tricks are helping to restock the antibacterial armoury.

BY NATASHA GILBERT

**T**he world is running out of effective antibiotics. Without action, bacterial infections that can now be shrugged off with a simple course of treatment could again become common causes of death. Antibiotic resistance in bacteria is blunting the effectiveness of drugs on which people have relied for almost a century, and too few new drugs have moved from the laboratory into clinical trials to add to the armamentarium. According to the World Health Organization, 51 antibiotics are in trials in people, but only 17 of those are considered to be innovative — with the remainder closely related to existing drugs. Fewer than 10 of the 51 are likely to make it through the minefield of drug development to market within five years.

More-responsible use of existing antibiotics will go some way to averting disaster. But the root of the problem is the historic neglect of research and development, says Suzanne Hill, director of the Department of Essential Medicines and Health Products at the World Health Organization. “Antibiotics are no longer of market value for pharmaceutical companies to research and develop,” she says. Instead, companies prefer to make more profitable investments in drugs that patients take over a long period of time. There is also a lack of basic research into the complex biology of bacteria, which has stalled the discovery of innovative drugs, she adds.

This call to arms is spurring researchers to search for new antibiotics. Here, *Nature* profiles some promising drug candidates and discoveries. Some comprise fresh twists on existing drugs or known targets long thought to have been overexploited. Others can better withstand attempts by bacteria to develop resistance, or are radical approaches — including a way to turn the bacterium’s immune system against itself. ■

A. B. DOWSETT/SPL

## RENEWED STRENGTH

The emergence of microbial resistance doesn’t have to spell ‘game over’ for an antibiotic. Entasis Therapeutics, a pharmaceutical firm in Waltham, Massachusetts, is working to revitalize the antibiotic cefpodoxime. This drug used to be a common line of attack against multidrug-resistant members of the Enterobacteriaceae family of microbes, which can cause serious infections in areas of the body such as the urinary and gastrointestinal tracts.

Cefpodoxime belongs to a group of broad-spectrum antibiotics known as  $\beta$ -lactams — named after the  $\beta$ -lactam ring in their chemical structures. Bacteria have evolved resistance to such drugs by producing enzymes called  $\beta$ -lactamases that break open the ring, destroying the drugs’ antibiotic properties. It’s a big problem — there are at least 3,000 types of  $\beta$ -lactamase, says Ruben Tommasi, chief scientific officer at Entasis.

Now, Entasis is turning the tables on resistant bacteria. The company is developing a compound called ETX1317 that binds to and inhibits  $\beta$ -lactamase, enabling  $\beta$ -lactam antibiotics to work uninhibited. Because ETX1317 has to be administered intravenously, it is best suited to use in hospitals to treat serious multidrug-resistant infections.

The company has also produced a version that can be taken orally, known as ETX0282. The World Health Organization has urgently called for new oral formulations of antibiotic, which will be of considerable benefit in the outpatient setting — for example, when treating people with urinary-tract infections that are complicated by resistant bacteria such as *Acinetobacter baumannii*. “This is a big deal. Doctors needed a go-to drug for urinary-tract infections,” says Robert Bonomo, who studies antibiotic resistance at Case Western Reserve University in Cleveland, Ohio.

Both Entasis compounds are effective against several multidrug-resistant Gram-negative bacteria, including *Klebsiella pneumoniae* and *Escherichia coli*, grown in culture, as well as their infections in mice. The company is conducting safety studies, and hopes to begin phase I clinical trials next year, says Tommasi.

Gram-negative bacteria are particularly difficult to beat because they have two cell membranes that drugs must traverse to be effective. By contrast, Gram-positive bacteria, including methicillin-resistant *Staphylococcus aureus*, have just one membrane, which makes them easier to penetrate. The World Health Organization has warned that “there is a serious lack of treatment options” for Gram-negative bacterial infections. Bonomo tips his hat to Entasis for taking on the challenge. “These are hard pathogens to treat,” he says.

The company is also looking to develop antibiotics to take the place of  $\beta$ -lactams, including a class that — like  $\beta$ -lactams — directly targets the penicillin-binding proteins that help to build bacterial cell walls. “All the  $\beta$ -lactam class of antibiotics are suffering some kind of emergence of resistance,” says Tommasi, but Entasis’s replacements are unaffected by the  $\beta$ -lactamases responsible. He is unable to reveal more about the new drugs, but the project has already attracted investors’ attention — winning up to US\$10.1 million from CARB-X, an international public-private partnership that funds preclinical antibiotic development. ■



## EVADING RESISTANCE

As quickly as researchers discover antibiotics, bacteria will evolve workarounds. But Kim Lewis, a microbiologist at Northeastern University in Boston, Massachusetts, is bullish that his candidate compound will stand up to the threat of resistance.

Many antibiotics bind to proteins inside bacteria. But pumps nestled in the bacterial cell wall can eject unwanted molecules from inside the cell. Lewis's antibiotic, teixobactin, fights microbes in a different way. It attaches to the outer surface of bacteria to avoid the ejection mechanism. Specifically, teixobactin binds to the molecular building blocks of two biopolymers — peptidoglycan and teichoic acid — that make up the bacterial cell wall. The compound acts to inhibit cell-wall synthesis.

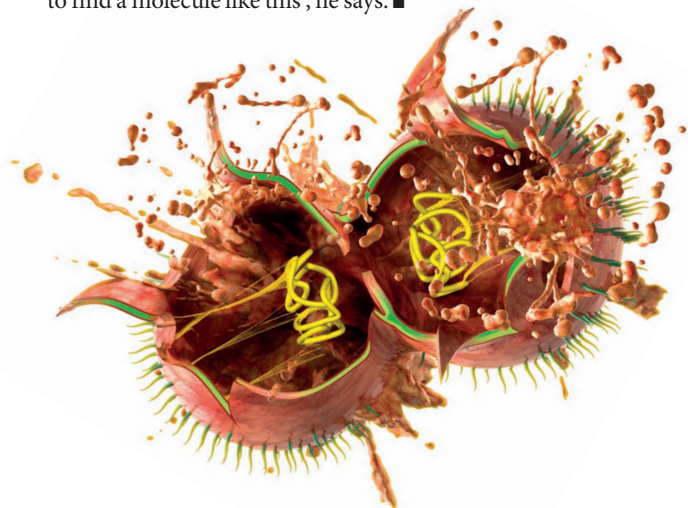
As well as binding to the outside of the cell, teixobactin has other advantages, says Lewis. The building blocks it targets are not directly encoded by DNA; rather, they are the product of a series of reactions catalysed by enzymes. This makes it less probable that bacteria will develop resistance because the extent of the changes required could not be achieved through a simple mutation alone. And as teixobactin binds to important regions of the building blocks, he adds, it is more probable that any mutation that did confer resistance would also adversely affect cell function, leading to a defective cell wall and the death of the bacterium.

Teixobactin is produced by *Eleftheria terrae*, a soil-dwelling species of bacterium that Lewis and his colleagues discovered using a nifty tool they developed. Because many microbes are difficult to cultivate on agar plates, the team invented the iChip — a thumb-sized device that holds hundreds of wells filled with a mixture of soil and agar that is diluted so that each well holds only one bacterium. The device is then planted in soil and the microbes grow successfully. “The bacteria are tricked into perceiving that they are growing in their natural environment,” says Lewis. Teixobactin is one of 30 potentially useful compounds that have been derived from the more than 10,000 microbes cultured using the iChip. Lewis is confident that further exploration of the soil microbiome using methods such as the iChip will yield important antibiotics. Researchers have identified only 1% of all microbes that exist, he estimates.

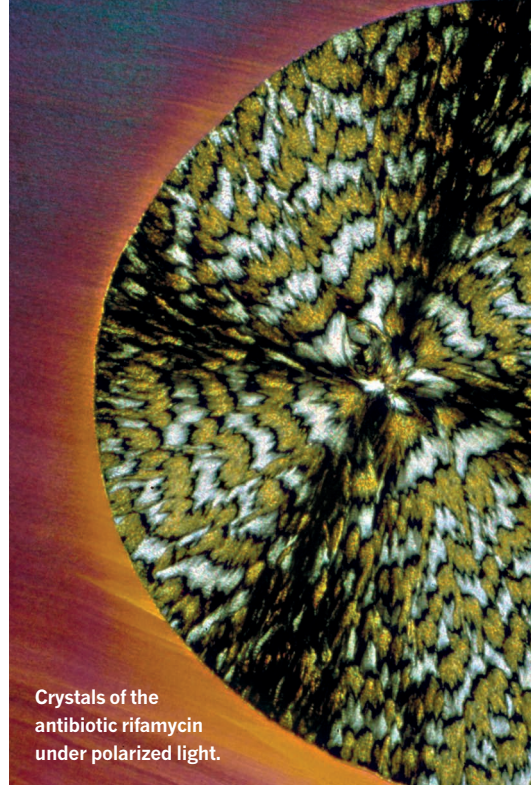
Tests in mice show that teixobactin is effective against the drug-resistant bacterium methicillin-resistant *Staphylococcus aureus*, as well as *Streptococcus pneumoniae*, which can cause pneumonia and meningitis. In culture, the compound defeated other disease-causing bacteria, including *Mycobacterium tuberculosis* and *Clostridium difficile*. Teixobactin will now undergo the testing required by the US Food and Drug Administration for researchers to get permission to start trials of the drug in people.

To examine how teixobactin holds up against the development of resistance, Lewis and his colleagues exposed *S. aureus* and *M. tuberculosis* to low doses of the compound. Such doses fail to kill all bacteria, and those that escape have the potential to evolve resistance. But no resistant microbes were found. “This indicates that teixobactin evolved to be largely protected against the development of resistance,” says Lewis.

The resistance-hardy compounds have buoyed researchers such as Timothy Lu, a synthetic biologist at the Massachusetts Institute of Technology in Cambridge, who is engineering antibiotics using the CRISPR–Cas9 gene-editing system. At a time when doctors are increasingly seeing antibiotic resistance, “it’s very exciting to find a molecule like this”, he says. ■



The antibiotic teixobactin (yellow) disrupts cell-wall formation in Gram-positive bacteria, leading to their rupture during cell division.



Crystals of the antibiotic rifamycin under polarized light.

DENNIS KUNKEL MICROSCOPY/SPL

## FRESH ATTACK ON A RESISTANT TARGET

Rifamycins are a group of front-line antibiotics that are used to treat infections such as tuberculosis and those that lead to pneumonia. These drugs help to kill bacteria by stopping the microbes from making RNA, a molecule that is essential for the production of protein, by inhibiting an enzyme called bacterial RNA polymerase. However, bacteria have evolved resistance to rifamycins by making a simple change to the amino-acid sequence of the RNA polymerase, which prevents such drugs from binding but does not impede the enzyme's ability to build RNA.

Scientists have discovered several other molecules that target bacterial RNA polymerase yet are different enough to evade the defences of bacteria. One such researcher, Richard Ebright, is close to being able to turn two of these molecules into potent antibiotics.

A molecular biologist at Rutgers University in Piscataway, New Jersey, Ebright has spent about two decades studying the structure of RNA polymerase in bacteria. He has been searching for uncharted binding sites on the enzyme, and then surveying bacteria that live in soil to see if any produce compounds that latch on to the sites. Although Ebright is bringing innovative techniques to the search, in many ways he is using an old-fashioned approach. “Our best source of new molecules has been microbial-extract screening,” he says. “Some people have said it is tapped out, but it’s not remotely tapped out. One simply needs to know how to look.”

Ebright has discovered six areas of interest on bacterial RNA polymerase. “The sites don’t overlap with the current drug-binding sites,” he says, meaning that any molecules that bind to them should be effective even in microbes that have already developed resistance to rifamycins. What’s

CLAUS LUNAU/SPL



more, these binding sites are common to the RNA polymerases of all bacteria, which makes Ebright's compounds good candidates for broad-spectrum antibiotics.

One of the sites is a hinge-like region that enables the RNA polymerase to open up, letting in DNA for translation into RNA. Ebright has found a compound called myxopyronin that stops the hinge from opening. Produced by the bacterium *Myxococcus fulvus*, myxopyronin has been used successfully to treat infections in mice. His team has since worked to improve the compound's potency and pharmacological properties, and myxopyronin is ready to enter clinical trials, says Ebright.

Another site is found in part of the enzyme that produces RNA from nucleotide building blocks. Ebright has found that a molecule called pseudouridimycin can take the place of a nucleotide, preventing the RNA polymerase from working. Pseudouridimycin has been shown to clear infection with *Streptococcus pyogenes* in mice. Ebright's team is now tweaking its chemical structure to increase the molecule's potency and stability.

Ebright hopes that, because these binding sites are in crucial areas of the RNA polymerase, this will prevent — or at least delay — the evolution of resistance to the new antibiotics. It will be harder for bacteria to alter such sites without affecting the activity of the enzyme. But he warns that, eventually, “bacteria will always find a way”.

Ebright's discoveries have made other researchers sit up and take notice. “The new chemistry is really exciting,” says Gerry Wright, who studies antibiotic resistance at McMaster University in Ontario, Canada. “It's a brand new chemical scaffold that hits the RNA enzyme in places that other drugs do not.” But the compounds have a long way to go before they prove themselves, Wright adds. “The difference between finding a new molecule and finding a new drug is huge.” ■

## PUSHING THE SELF-DESTRUCT BUTTON

Many bacteria defend themselves against invading viruses through an immune system called CRISPR — more widely recognized in the past few years for its application in genome editing. After a bacterium has been exposed to a virus, also known as a bacteriophage, its CRISPR system generates a short RNA sequence that is complementary to a specific part of the phage's genetic code. When the bacterium is infected again, the RNA can then guide enzymes to cut the phage's DNA — often destroying the virus.

Locus Biosciences, a biotechnology company in Research Triangle Park, North Carolina, aims to subvert this CRISPR system. “We're able to harness and activate the bacterium's natural immune system to kill itself,” says Paul Garofolo, the company's co-founder and chief executive.

Researchers at Locus are arming phages by loading them with DNA that matches sequences found in the bacterial genome. The viruses can then infect bacteria and insert their genetic material into the nucleus. When the viral DNA is transcribed, the resulting RNA guides the CRISPR system's cutting enzyme to several targets in the bacterial genome. However, unlike CRISPR-mediated genome editing, which uses the enzyme Cas9 to make clean cuts to both strands of DNA, the Locus system uses Cas3. “Cas3 doesn't just cut DNA, it degrades it at the same time, so it can't be repaired,” says Dave Ousterout, co-founder and chief technology officer at Locus.

Timothy Lu, a synthetic biologist at the Massachusetts Institute of Technology in Cambridge, has developed a method of targeting bacteria that is based on the CRISPR–Cas9 system. He says that both the Cas3 and Cas9 enzymes are “useful ways to trigger intracellular DNA cutting”, and that “both of these strategies may be employed to kill”.

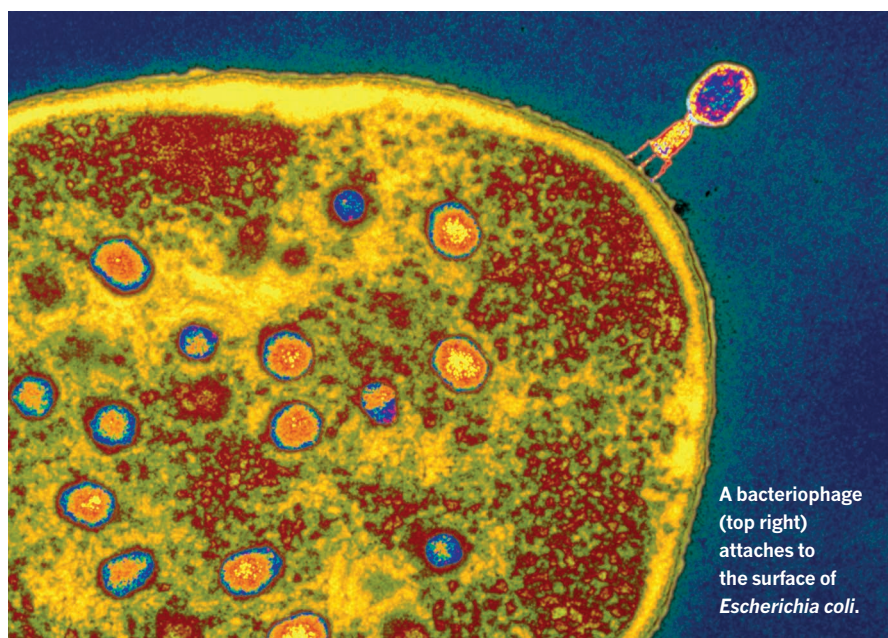
Locus is mainly focused on tackling intestinal pathogenic agents such as *Clostridium difficile* and *Escherichia coli*. Antibiotic-resistant *C. difficile* poses one of the most urgent threats to human health, and *E. coli* can cause life-threatening infections of the blood and urinary tract. In laboratory tests, the CRISPR–Cas3 antibacterial tool cleared *C. difficile* infections in mice, says Garofolo. The company hopes to start phase I trials in 2019, but must first obtain approval from the US Food and Drug Administration. Garofolo adds that the treatment would probably be administered to people who do not respond to existing first-line drugs such as vancomycin.

Countries such as Russia and Poland have long used phages to treat bacterial infections in people. The advantage of this treatment is that it targets only specific bacteria, rather than wiping out a swathe of beneficial bacteria along the way. But phage therapy hasn't taken off more widely, in part because bacteria can easily develop resistance to the phages.

Garofolo hopes to see higher efficacy with the CRISPR–Cas 3 system than with conventional phage therapy because the viruses used by Locus have been boosted with bacterial DNA. Furthermore, the team hopes to limit the risk of bacteria developing resistance by using multiple phages to attack the bacterial genome at several sites — ensuring that the bacteria cannot survive. And restricting the use of the CRISPR–Cas3 treatment to people with the most severe infections will also help to limit the opportunities for resistance to develop. “We anticipate that some sort of antibiotic stewardship upfront will help our product maintain efficacy,” says Ousterout.

The phage therapy developed by Locus has broad potential, says Garofolo. The company's goal is to use the technology to treat long-term conditions such as irritable bowel syndrome or colorectal cancer. “If the technology breaks through,” he says, “it should be able to address any number of additional bacterial targets.” ■

**Natasha Gilbert** is a freelance science writer based in Washington, DC.



# Teaching tolerance

*Researchers are seeking to tame the immune system with the aim of alleviating, or even preventing, autoimmune disease.*

BY KATHERINE BOURZAC

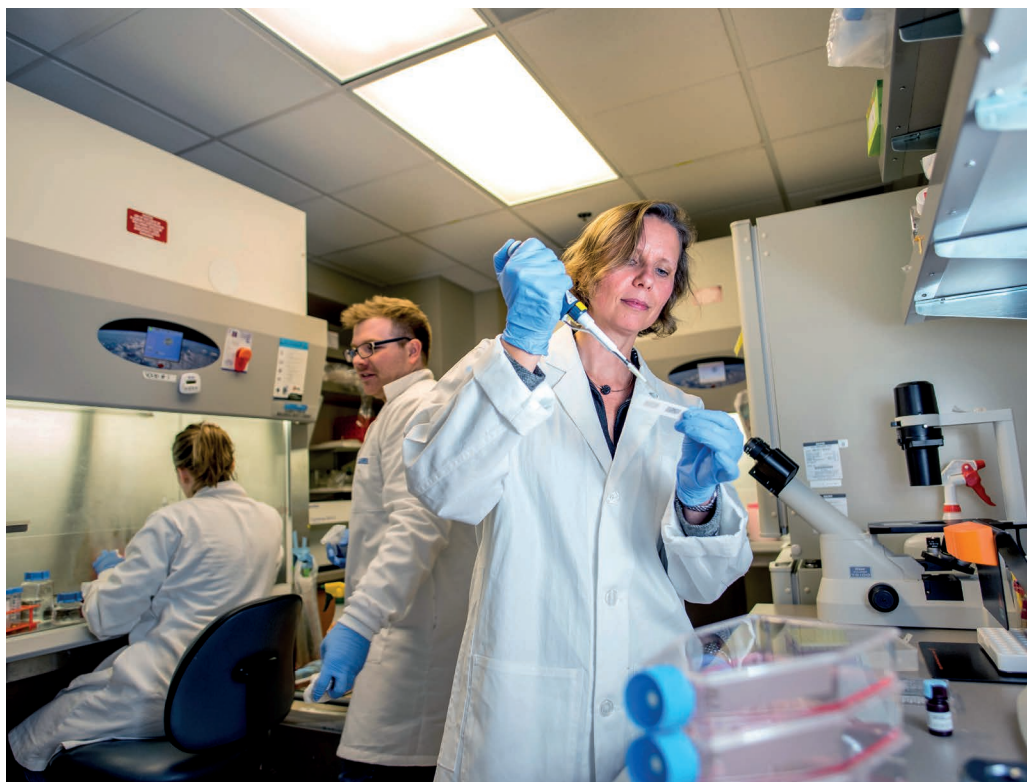
There is no cure for type 1 diabetes. Instead, the condition, in which a person's immune system destroys their ability to produce the glucose-regulating hormone insulin, must be carefully managed. But even the most vigilant of those affected will experience organ-damaging bouts of high blood-sugar levels, and will be at an increased risk of cardiovascular disease, nerve damage and blindness.

Immunologists suggest that it doesn't have to be that way. In August 2017, Mark Peakman, a clinical immunologist at King's College London, and his colleagues published the results of an early clinical trial of a treatment that aims to teach the immune systems of people with type 1 diabetes to spare the insulin-producing cells of the pancreas<sup>1</sup>. The study was designed to assess the safety of the treatment in those who have been newly diagnosed with the disease, but it also showed hints of efficacy, says Peakman. Six months after their initial diagnosis, most of the treated participants were still producing enough insulin to avoid the need to increase their use of synthetic insulin, unlike those who received a placebo. The researchers are now planning a phase II clinical trial.

"In principle, you would treat this as early as possible, in people who have high-risk genes," says Peakman. He calls the approach "extreme prevention".

Many researchers are setting up clinical trials of treatments for type 1 diabetes and other incurable autoimmune diseases, such as multiple sclerosis and Graves' disease, to test ways of bringing hyperactive immune systems into line. Immunologists think that by treating people with molecules that can induce an immune response (antigens), bacteria or engineered immune cells, it could be possible to train the immune system to tolerate the tissue it is on track to damage — an intervention that has the potential to cure a range of autoimmune disorders.

There is much to prove. So far, this new generation of treatments has proved to be safe, but its efficacy is uncertain — the field is one in which the findings of clinical trials often fail to reflect encouraging results from the laboratory. But armed with a deepening understanding of the molecular basis of autoimmunity, as well as advances in genetic engineering and cell-based



Megan Levings and her team are engineering regulatory T cells to help keep autoimmune diseases in check.

therapy, immunologists are hopeful that, this time, the results will be different.

## GROWING PROBLEM

Autoimmune diseases can affect almost any part of the body. In each case, the body loses tolerance towards its own tissues — certain proteins are seen as antigens, and the immune system attacks. Lack of tolerance also jeopardizes recipients of organ or bone-marrow transplants, leading to rejection of the transplanted organ or causing immune cells in the bone-marrow graft to attack the recipient's body.

Autoimmune conditions are on the rise, particularly in the developed world and in women<sup>2</sup>. "There are numerous environmental influences," says David Wraith, an immunologist at the University of Birmingham in the United Kingdom. Precisely what such influences are is unclear, but suggestions include diet, exposure to sunlight and pollution, and stress, he says.

Researchers also anticipate that the growing number of people with cancer who are treated with immunotherapy will further increase the

prevalence of autoimmunity. Such treatments deliberately unleash the immune system to fight tumours, but they can also trigger autoimmune diseases, including rheumatoid arthritis and colitis. In May 2017, researchers at the University of California, San Francisco, and the nearby Parker Institute for Cancer Immunotherapy, proposed that these unintended consequences have become "the Achilles' heel of immunotherapy"<sup>3</sup>.

Conventional treatments for autoimmune diseases are limited either to managing symptoms or to aggressively targeting the whole immune system, which can cause side effects and make recipients vulnerable to infection. Immunologists see a third way: downregulation of only the specific immune reactions that are harmful.

"We know what we want the immune system to do, but it's been really hard to do it," says Megan Levings, an immunologist at the British Columbia Children's Hospital Research Institute in Vancouver, Canada. "The key is to have an antigen to target."

Tapping into the immune response's existing

PAUL JOSEPH



control system, and doing it with specificity, is seen as the way forward. The main immune players are the regulatory T ( $T_{reg}$ ) cells, which Levings calls “the brakes of the immune system”. Even when it is responding to genuine infections, the immune system can go too far, causing harmful inflammation.  $T_{reg}$  cells help to prevent this. Similarly to other T cells, they are activated by specific antigens. But instead of mounting an attack,  $T_{reg}$  cells rein in the immune cells that are doing damage.

Wraith and other immunologists think that the body can be made to produce the  $T_{reg}$  cells required to dampen a certain autoimmune response, by dosing people who are affected with the same antigen or antigens that the immune system wrongly interprets as a reason to attack. The approach is counter-intuitive: antigens such as those given in a vaccine typically put the immune system on alert. But if administered without the immune-system stimulants called adjuvants that are usually included in vaccine formulations, antigens can induce a calming effect through  $T_{reg}$  cells.

Wraith's group completed a phase II trial in people with relapsing multiple sclerosis in 2016. Those with the disease develop lesions in the brain and spinal cord when the immune system attacks the protective sheath that surrounds nerves. Wraith's experimental treatment, which has been licensed by Apiteo, a biotechnology company based in Chepstow, United Kingdom, comprises a cocktail of peptides from especially antigenic regions of myelin basic protein — the main target of the immune system in multiple sclerosis. Wraith says that the people they treated had less inflammation in their brains, as measured by magnetic resonance imaging.

Peakman's trial for a type 1 diabetes treatment used a single peptide based on proinsulin, a precursor to insulin and the antigen that the immune system targets in the pancreas. But similar to Wraith, he and his colleagues intend to use a cocktail of peptides from the targeted protein in later phases of the study. Although loss of tolerance to a particular protein such as proinsulin lies at the core of many autoimmune diseases, immunologists think that other antigens could also contribute to the manifestation of such diseases in some individuals — making a cocktail of peptides more likely to succeed than one alone. In Peakman's preclinical studies, this approach has been more effective than using a single peptide. “More peptides are better,” says Peakman.

### CELL-BASED THERAPY

There may be other ways to tame a rogue immune system. Researchers propose that bacteria dwelling in the body thrive by inducing immune tolerance, and would like to turn this to the advantage of medicine. In February 2017, researchers led by endocrinologist Chantal Mathieu at the Catholic University of Leuven in Belgium reported that genetically

modified *Lactococcus lactis* bacteria can reverse diabetes in two-thirds of mice with the condition by inducing  $T_{reg}$  cells<sup>4</sup>. The bacteria were engineered to produce proinsulin and an anti-inflammatory cell-signalling molecule called interleukin-10. This work has been licensed by Intrexon Actobiotics of Gent, Belgium, and will enter clinical trials this year, says Mathieu.

Levings sees promise in manipulating the  $T_{reg}$  cells of patients more directly, by removing them from the body, training them and then returning them. She is working on ways to engineer large numbers of  $T_{reg}$  cells to respond to specific antigens that have been wrongly recognized by an individual's immune system as being foreign. Levings' lab modifies T cells using a protein called a chimaeric antigen receptor (CAR) — a method that has already been approved for use in cancer treatment. Whereas cancer researchers use CAR proteins to make T cells attack tumour cells, Levings uses them to make  $T_{reg}$  cells that will dampen harmful inflammation.

She is able to engineer the cells to respond to an antigen of choice. At the moment, her lab is focusing on type 1 diabetes. Levings sees the potential to prevent a lifetime of complications in young people who have just been diagnosed with the disease. Such a treatment would be expensive, but could transform the quality of life of its recipients. “When I started, people looked at me like I was crazy,” says Levings. Because a low-risk treatment for type 1 diabetes already exists — the administration of synthetic insulin — there has been little acceptance so far of the potential risks of genetically engineering a person's immune cells. That's changing, thanks to a growing safety record for the method in treating cancer. “Cancer immunology has changed the world's perception of cell therapies,” says Levings.

$T_{reg}$ -cell therapy might also offer a way to induce tolerance following an organ or bone-marrow transplant. When a solid organ such as a kidney is transplanted, the recipient can have an immune reaction to the donor tissue.  $T_{reg}$ -cell therapy could stop the reaction without systemically weakening the immune system and leaving the recipient vulnerable to infection. The transplanted tissue can also be the source of autoimmunity — in graft-versus-host disease (GVHD), immune cells from the grafted tissue attack the recipient's body. The standard therapy for GVHD is corticosteroids, but about half of the people treated do not respond, says Bruce Blazar, a clinician and researcher at the Pediatric Bone and Marrow Transplant Center of the University of Minnesota in Minneapolis. Reining in GVHD is particularly challenging. Whereas autoimmune diseases are typically confined to a single tissue, transplanted immune cells can go anywhere. “After a bone-marrow transplant, the entire body can be subject to GVHD,” Blazar says.

In mice, infusions of  $T_{reg}$  cells seem to

help. Robert Zeiser, a clinical oncologist and haematologist at the University of Freiburg in Germany, says that the effects of treating GVHD with engineered  $T_{reg}$  cells in preclinical studies are dramatic. “The GVHD mice were terribly sick, but the group treated with  $T_{reg}$  cells looked completely healthy,” he says. “I have not seen any other approach that can block GVHD so effectively.”

In people, however, it is unclear whether the antigen-specific approach to tackling autoimmune disease can prevent GVHD. For instance, researchers often struggle to narrow down the molecules that the grafted cells will perceive as antigens. Studies in mice suggest that

50 to 100 antigens are typically involved in GVHD, but “we don't know how many antigens are important in people”, Blazar says.

The possible antigens behind the immune reaction involved in transplant rejection are less numerous and therefore easier to identify. Flavio Vincenti, a specialist in kidney and pancreas transplants at the University of California, San Francisco, hopes that  $T_{reg}$  cells can help to prevent organ rejection in people showing signs of inflammation after a kidney transplant. His group is recruiting participants for a phase II trial that will compare the efficacy of infusing transplant recipients with a population of their own  $T_{reg}$  cells that has been trained to recognize antigens in the blood of the kidney donor, or with an expanded general population of their own  $T_{reg}$  cells.

Levings is excited about the rejection trials. “This is the perfect clinical context to test the cell therapy,” she says. “You know what the antigens are, you know what the mismatch is between donor and recipient, and you have control over the day it's going to be done.”

Vincenti's expectations, however, are more measured. Getting to this point has been difficult because the immunosuppressive drugs used to prevent transplant rejection work so well, he says. “We are a victim of our own success.” Innovative treatments such as infusions of  $T_{reg}$  cells have the potential to prevent side effects in the long term, but the upfront risks are greater.

However, he is excited to learn from the trial. “If we don't make the first step now, how are we going to make the giant step to new therapies?” he says. “We may fail, but we will learn a lot.” ■

**Katherine Bourzac** is a freelance journalist in San Francisco, California.

1. Alhadi Ali, M. et al. *Sci. Transl. Med.* **9**, eaaf7779 (2017).
2. Wraith, D. C. *Front. Immunol.* **8**, 1668 (2017).
3. June, C. H., Warshawer, J. T. & Bluestone, J. A. *Nature Med.* **23**, 540–547 (2017).
4. Takiishi, T. et al. *Diabetes* **66**, 448–459 (2017).

## PHARMACEUTICALS

# A CRISPR path to drug discovery

*Gene editing is quietly revolutionizing the search for new drugs.*

BY ANDREW SCOTT

CRISPR–Cas gene editing, once considered arcane, is fast entering mainstream use in research. Most people with an interest in science have probably heard about the technique, which uses a combination of a synthetic guide RNA molecule and an enzyme (typically Cas9) from the bacterial immune system to edit DNA with unprecedented ease and precision. It is a flexible tool with a variety of applications. Most of the media interest in the CRISPR–Cas system has focused on its potential for treating diseases with a genetic basis. Yet CRISPR–Cas also has a big part to play in drug discovery, which could prove to be as important as its therapeutic use — if not more so.

In a comprehensive 2017 review, scientists from the University of California, Berkeley, including co-discoverer of CRISPR–Cas Jennifer Doudna, emphatically concluded that this type of gene editing is “ready to have an immediate impact in real-world drug discovery and development”<sup>1</sup>.

Christof Fellmann, a biotechnologist and co-author of the review, explains that the ability of CRISPR–Cas to help identify target molecules will have a crucial impact on drug discovery. By using the system to deliberately activate or inhibit genes, researchers can determine the genes and proteins that cause or prevent disease, therefore identifying targets for potential drugs. CRISPR–Cas is also making it easier to create cellular and whole-animal model systems that precisely mimic diseases. This is enabling scientists to more accurately verify the safety and efficacy of drugs, which ensures that such models are better predictors of what will happen in clinical trials. As these uses are pursued, researchers are also refining and extending the capabilities of CRISPR–Cas to make it an even more powerful gene-editing tool.

“It makes everything easier,” says Jon Moore, chief scientific officer at biotechnology company Horizon Discovery in Waterbeach, near Cambridge, United Kingdom. In March 2016, at an event at the Science Museum in London, Moore declared, “The targets we’re finding with CRISPR–Cas9 are going to guide the

drugs coming out in the 2020s.” He stands by that assessment two years on. “If it’s not right, then I’ll be in trouble,” he laughs.

## A KNOCK-OUT TOOL

The mechanism that underlies CRISPR–Cas gene editing is relatively simple. A short strand of RNA, tailored to target a specific sequence of DNA, is linked to an enzyme that is capable of cutting double-stranded DNA. Cas9, the enzyme to which Moore refers, is the most widely used, but other enzymes are being explored. After the RNA and enzyme are delivered to the cell nucleus, the RNA binds to its complementary DNA sequence, acting as a guide for the enzyme that then chops the DNA. After that crucial cut is made, DNA-repair enzymes in the cell fix the break in a way that either disables or modifies the targeted gene — its activity can be turned up or down, mutations can be introduced, or sections can be inverted.

The simplicity of using guide RNAs to target any location in the genome is making gene editing accessible to many more researchers. “CRISPR–Cas has taken gene editing out of the hands of those specialists who are expert in complicated molecular biology,” says Moore.

Researchers engaged in drug discovery are eagerly exploiting CRISPR–Cas to switch off — or ‘knock out’ — specific genes to see what they do. Methods of introducing such knock-out mutations have been in use since about 2000, but these earlier approaches, which rely on engineered enzymes to cut DNA, often only partially knock out genes, commonly produce unwanted effects on unintended targets, and lead to inconsistent results between similar studies. CRISPR–Cas avoids these deficiencies and, since rising to prominence in 2012, has made it straightforward to knock out genes of choice. “The difference is in the quality of information you can get,” says Moore. CRISPR–Cas is better at knocking out the targeted gene more fully, as well as avoiding unwanted effects, which has made large-scale gene-function experiments much more reliable, he explains.

Knock-out screening to identify genes involved in drug resistance is fast becoming



one of the most widely used applications of CRISPR–Cas gene editing in drug discovery. Researchers expose large numbers of cells to a pool of CRISPR–Cas systems carrying guide RNAs that target various genes. This allows them to generate and select individual cells that each have a specific gene knocked out. The cells are then exposed to chemicals or drugs of interest. Genes that confer resistance to drugs can be identified through cells that become sensitive to such compounds after the CRISPR–Cas treatment. These genes, or the proteins they encode, can then be targeted with other drugs to get around the problem of resistance.

Identifying genes that promote disease uncovers some obvious targets for drug development. The simplest candidate drugs bind to and interfere with the proteins encoded by these genes, rather than affect the genes directly. But more-subtle targets for drugs can be revealed by a better understanding of

the importance of multiple genes and proteins, their interactions and their mutual regulatory effects. Many diseases, for example, arise when things go wrong in a regulatory pathway that involves a complex network of intracellular interactions. Using CRISPR–Cas to identify, with ease and accuracy, combinations of genes involved in these networks should offer a more sophisticated approach to treatment.

Researchers are also using CRISPR–Cas and the DNA-repair processes of the cell to incorporate — or ‘knock in’ — selected sections of DNA. This can introduce mutations

MICHELE MARCONI





Berkeley, has added a binding site for the hormone oestrogen to the enzyme Cas9 (ref. 3). This demonstrates the possibility of subtly controlling the activity of the gene-editing system through an external signal such as the level of a hormone or its analogue. As an example of its utility in drug development, Fellmann says that the technique could be used to control the timing of gene editing to more closely mimic the timing of the effects of drug molecules in disease models. These engineering efforts are at an early stage, but should eventually lead to a range of innovative functions.

Drug development is a long process: it can take more than a decade for researchers to move from the discovery of a target molecule to the production of a clinically approved drug. So it could be some time before the first drugs to be developed using CRISPR–Cas gene editing hit the market. “But,” says Fellmann, “people are already using it now, and in the long term it will definitely have a significant impact.”

Jonathan Wrigley, associate director of the Innovative Medicines and Early Development Biotech Unit at AstraZeneca in Cambridge, offers a similarly confident outlook from big pharma. “We are applying CRISPR–Cas technology across our drug-discovery pipeline,” says Wrigley. He says that teams at AstraZeneca have generated more than 100 disease models with the aid of CRISPR–Cas in the past three years, and are constantly finding ways to improve the technology. “It has proved transformative in the generation of cellular models to support drug-discovery projects,” Wrigley adds. Engineered cell-based models with precise genetic modifications were previously rare in drug discovery, owing to the challenging and time-consuming techniques that were required to generate them. “The CRISPR–Cas technology has enhanced both the feasibility and speed of this process, thereby enabling such models to become integral tools in the early stages of our drug-discovery projects, in a manner not seen before,” Wrigley says.

His is one of countless research groups throughout academia and the pharmaceutical and biotechnology industries that now use CRISPR–Cas tools in the search for drugs. Predicting when gene editing will bear fruit is difficult; the drug-development pathway is long and clinical trials are laced with uncertainty, no matter the tool. But with so much ongoing activity, Moore’s prediction that CRISPR–Cas will transform drug discovery seems unlikely to get him into trouble. “There’s been a massive investment in CRISPR–Cas by pharma,” he says. “I am not alone.” ■

**Andrew Scott** is a science writer in Perth, UK.

1. Fellmann, C., Gowen, B. G., Lin, P. C., Doudna, J. A. & Corn, J. E. *Nature Rev. Drug Discov.* **16**, 89–100 (2017).
2. Murovec, J., Pirc, Ž. & Yang, B. *Plant Biotechnol. J.* **15**, 917–926 (2017).
3. Oakes, B. L. et al. *Nature Biotechnol.* **34**, 646–651 (2016).

that transform the protein encoded by the targeted gene, leading to beneficial effects that a drug could then be designed to induce more simply. Some variants of CRISPR–Cas systems can make changes that either inhibit or promote the activity of a gene without changing its actual function. Turning gene activity up or down is a subtler way of investigating the importance of genes and proteins that could be activated or inhibited by drugs to treat disease.

“CRISPR–Cas is enabling nearly unlimited genetic manipulation,” says Fellmann, and it is bringing researchers much success. “We have already found exciting new targets using CRISPR–Cas technology,” says Moore. He will not reveal what the target molecules are, but does say that his company’s research involves mutations in “undruggable” tumour-suppressor and cancer-causing genes that other researchers have been unable to target.

### MODEL MAKING

Cell and animal models of human disease are crucial elements of drug development. The initial stages of testing candidate drugs for efficacy and toxicity can rarely be done in people, for ethical reasons. However, many of the disease models that are available to researchers are far from perfect. The main problem has been the complexity — and therefore the time and expense — of building superior models for the huge variety of human diseases that exist. “In industry, speed and cost are as important as feasibility,” says Fellmann. If it would take too long and cost too much to make a great model, a less perfect one might be preferred. Yet the developers of drugs would like to avoid such a compromise.

Both Moore and Fellmann agree that the simpler and more reliable gene editing made possible by CRISPR–Cas has enabled researchers to create models of disease more quickly and cheaply. “We can now, pretty much, change any gene in whatever way we want to mimic a disease,” says Fellmann. He also emphasizes that the “surgical precision” of CRISPR–Cas gene editing means that little or no trace remains of the editing process. With older genetic-engineering techniques, extra changes to the DNA sequence can be left in or around the altered genes, similar to a surgeon leaving instruments inside a patient after an operation. The precision of CRISPR–Cas greatly reduces the chances of the gene-editing tool having an undesired effect.

### WAYS TO IMPROVE

Fellmann and his colleagues are now trying to find and develop innovative versions of the existing CRISPR–Cas tools that might bring further flexibility and precision. Part of this effort is exploring other bacterial gene-editing systems, and alternatives to Cas enzymes have already been found<sup>2</sup> (see *Nature* **536**, 136–137; 2016). An enzyme known as Cpf1, for example, can cut DNA at sites to which CRISPR–Cas is unable to bind. Other such enzymes, including Cas13, can cut RNA — the intermediary between DNA and protein — rather than DNA. This opens up flexibility in the options for modifying the activity of genes, beyond their basic editing.

A more adventurous approach is to engineer the genes from bacteria that encode the enzymes used in existing CRISPR–Cas systems, to add extra abilities. For example, a team of researchers at the University of California,



## BIOENGINEERING

# The power of thought

*Neural prostheses are helping to restore movement and the sense of touch in people with paralysis.*

BY NEIL SAVAGE

Eugene Alford just couldn't get his legs to move, but it wasn't for want of trying. It was 2012, and he was in a laboratory at the University of Houston in Texas, participating in a study that was designed to see whether people with paralysis could control a robotic exoskeleton with their thoughts. Alford, a plastic surgeon who'd lost the use of his legs when a tree fell on him at his farm, kept trying to walk by willing the electrical impulses in his brain up and into the electrodes on his head, from where they could be translated into movement.

Jose Contreras-Vidal, the neural engineer who was conducting the experiment, urged Alford not to think too specifically about the act of walking. Instead, he should just concentrate on where he wanted to go. "Finally, he put a cup of coffee on the desk, and I started thinking, 'I want that cup of coffee,'" Alford, now 58, says. So Alford strode over to the desk and took it. By thinking about walking as an able-bodied person would — that is, by barely thinking about it at all — he was able to send the correct signals to the brain-machine interface that controlled the robot.

The movement that the technology bestowed was a big deal for Alford. "Just being able to stand up and look somebody face to face, in the eye, for a person who's been in a wheelchair for five years, that's what brings tears to your eye," he says. Six years on, Contreras-Vidal's lab at the Building Reliable Advances and Innovation in Neurotechnology Center, a collaboration between the University of Houston and Arizona State University, continues to train paralysed people to walk, albeit only under the supervision of researchers. His group is one of a number that are developing practical

**An exoskeleton controlled by brain activity is tested by an able-bodied boy.**





neural prostheses — devices capable of reading signals from the brain and then using them to restore movement in people who have been paralysed through injury or illness.

The World Health Organization estimates that 250,000–500,000 people worldwide suffer a spinal-cord injury every year, about 13% of whom will lose the ability to control all four limbs. Another 45% will retain some movement or feeling in all limbs, but are still severely limited in what they can do physically. And almost 2 million people affected by stroke in the United States are living with some degree of paralysis, as are another 1.5 million people with multiple sclerosis or cerebral palsy.

Against this backdrop of paralysis, researchers are working to engineer technological solutions. As well as enabling the control of robotic aids, some groups are learning to detect the brain's intention to initiate movement and to then feed that instruction into the muscles. A few groups are also trying to send signals back into the brain to restore sensation in people who can no longer feel their limbs. But before these technologies can touch lives beyond the lab, researchers must improve the understanding of how best to integrate humans with machines.

## **A CLOSER LISTEN**

Contreras-Vidal records electrical activity in the brains of his study volunteers through a skull cap that is studded with 64 electrodes. The impulses gathered are then translated into signals to control the robotic exoskeletons.

Listening to populations of neurons using electrodes mounted outside the skull is not a simple task. Like hearing music from across the street, some subtleties are lost. And movement of the scalp muscles, eye blinking and motion in the wires that connect the electrodes to the decoder all add noise that makes the neural signals trickier to interpret. The system provides enough information to unravel the user's intentions and to translate them into movement, but other researchers are using implanted electrodes to read signals from individual neurons, in the hope of collecting a more nuanced signal and providing finer-grained motor control.

In 2016, Bill Kochevar of Cleveland, Ohio, became the first person with paralysis to use electrodes implanted in the motor cortex of his brain to stimulate his arm to move. Implanted electrodes had already enabled people with a spinal-cord injury to move robotic arms, but, thanks to a combination of the brain implants and a set of stimulatory electrodes in his right arm, he was able to move his arm to feed himself, raise a cup to his mouth and scratch his nose. Although these regained abilities were limited, they still opened up his world. "I know there are a lot more possibilities out there for doing things I didn't think were possible," he said in October last year. "It's always been exciting to me that I'm first in the world to do this."

The feat brings doctors closer to restoring lost

function in people with paralysis. "It's a big deal scientifically, but it's also a big deal clinically," says Bolu Ajiboye, a biomedical engineer at Case Western Reserve University in Cleveland, who worked with Kochevar. "He couldn't do anything on his own before."

Kochevar was in his mid-forties when he crashed his bicycle into the back of a postal truck, injuring the top of his spine, and causing him to completely lose the ability to move his limbs. He died last December at the age of 56 from complications of that injury, having participated in the implant research for about three years. The researchers he assisted — part of BrainGate, a collaboration between Case Western, Brown University in Providence, Rhode Island, Massachusetts General Hospital in Boston, Stanford University in California, and the US Department of Veterans Affairs — are continuing to recruit volunteers.

To enable Kochevar to move his arm, the researchers implanted two square arrays of 100 electrodes, both 4 millimetres long, in the area of his motor cortex that was responsible for hand movement. Another 36 electrodes implanted under the skin of his right arm provided tiny jolts to the muscles in his hand, elbow and shoulder through a technique known as functional electrical stimulation. The brain arrays were wired to bolt-like connectors that protruded from the top of his head. Cables carried signals from the connectors to a computer, which applied machine-learning to the data to ascertain the movements that Kochevar wanted to make. The electrodes in his arm then received a pattern of stimuli that caused his muscles to move. Because Kochevar's muscles had weakened through disuse, the researchers also provided him with a motorized arm support, which received the same movement commands as his own muscles.

Before Kochevar could start to use the system, the researchers had to train the computer to interpret his intentions. Initially, they asked him to watch a moving arm in virtual reality while imagining that he was making the same movements. Later, they tried a lower-tech approach that Ajiboye says worked just as well; they moved Kochevar's arm using the computer and had him imagine he was doing it.

The imagined movements created distinct patterns of activity in the 200 or so neurons in Kochevar's brain that were being monitored individually by the two implants. The researchers recorded the order and rate of neuron firing for each movement, enabling them to stimulate the correct muscles in Kochevar's arm when a particular pattern of movement was detected in subsequent experiments.

To begin with, Kochevar had to concentrate on the individual movements that comprise a gesture. "When I first started doing it, I thought a lot about moving in, out, up, down," he said. But as time went on, he was able to go beyond purely mechanical directives. With practice, moving his arm came more naturally;

like Alford, he learnt to think about what he wanted to do rather than how to do it. "I just think about going from here to there, and it pretty much goes there," he said.

He only ever used the system under the supervision of the researchers, either in the lab or at his home, owing to the complexity of the set-up and US Food and Drug Administration (FDA) safety regulations. Ajiboye and his colleagues needed to calibrate the system at the start of each day of testing, to ensure that the electrodes were aligned correctly in the brain. Although the day-to-day drift is usually small, in time the implants could end up recording a different group of neurons, which would mean having to interpret a fresh set of activity patterns. Calibration takes around five minutes, but Ajiboye hopes that his team will eventually reduce it to just a few seconds.

## **AN UNCOMMON TOUCH**

The only feedback that participants in Ajiboye and Contreras-Vidal's research receive when they move is visual. Other researchers, however, are trying to provide users with another type of important sensory information — touch. "That's how we know to hold objects the right way, to make sure we don't crush them or that they don't fall out of our grasp," says Robert Gaunt, a biomedical engineer at the Rehab Neural Engineering Labs at the University of Pittsburgh in Pennsylvania. Sight alone does not always provide enough information for a person to judge whether he or she is touching an object, or to guess the correct firmness of grip, and physical sensation is crucial to the fine-grained control that is required to write with a pen or to turn a key.

In 2015, Gaunt and his colleagues began to test such a feedback system in Nathan Copeland, a 28-year-old man from Pennsylvania who had been paralysed in all four limbs in a car accident a decade earlier. Like

Kochevar, Copeland had electrodes placed in his motor cortex as part of a separate experiment led by Gaunt to control a robotic arm. To provide Copeland with a sense of touch, however, the researchers needed to implant two arrays of electrodes in his primary somatosensory cortex — the area of the brain that is responsible for registering such sensations. These arrays were wired to pressure sensors in the hand of the robotic arm, and the researchers pressed each finger separately, out of Copeland's sight. He correctly identified which they were touching 84% of the time, and was usually right about the index and little fingers, but occasionally mixed up the middle and ring fingers.

The sensation wasn't entirely the same as touch. Often, Copeland did describe the feeling as touch or pressure, but some of the electrodes produced other sensations,

*"I just think about going from here to there, and it pretty much goes there."*



Nathan Copeland (background) can experience touch through a pairing of brain implants with a robotic hand.

including tingling, buzzing or warmth. The researchers are trying to understand what causes these responses, in the hope that they can use them to the advantage of people with paralysis. It might be useful, for instance, for them to be able to feel temperature.

Finding appropriate sensors for the various types of sensory information won't be a problem; sensors have been built for tasks as varied as controlling industrial robots and providing haptic feedback in smartphones. In Montrose, California, robotics company SynTouch has even developed a way of distinguishing one texture from another. But neurologists are not yet ready to take advantage of such possibilities, Gaunt says. "We still have no idea of how to send that sort of information into the brain."

#### SMALLER AND SOFTER

Neural prostheses are a long way from being ready to use in a domestic setting. One problem is that they're bulky and obtrusive. "The system is essentially a rack of computers that record the brain activity," says Ajiboye. "We need to miniaturize the recording technology so it's the size of a cell phone and it can sit on the side of a wheelchair." He'd also like to use wireless sensors to eliminate the need for users to be tethered physically to a computer.

Researchers are already working on such sensors, and groups have tested several in rats and monkeys in the past five years. But wireless technology alone will not clear the path to widespread use. The biggest impediment to implanted electrodes is that they tend not to last for more than a few years. The only

electrode system that has FDA approval for implantation in the human brain is the silicon-based Utah array — the device type placed in both Kochevar and Copeland. Each device's needle-like electrodes are 0.5–1.5 millimetres long and stick into the brain. Developed in the 1990s, the array provokes an immune response that produces a local build-up of tissue called a glial scar, which limits the flow of electrical signals. And because the device is much harder than brain tissue, it can drift out of alignment as a person moves, changing which neurons it records and promoting further tissue irritation.

Although researchers have managed to get useful information from the arrays for five or more years after implantation, the signals obtained become progressively less detailed. Even in relatively young devices, some areas will have stopped working, says Jeffrey Capadona, a biochemist and materials scientist at Case Western who studies why the arrays fail. To be practical for use in people with a spinal-cord injury, who can live for decades after the initial trauma, the implants would have to last much longer. An array that could span a person's lifetime would remove the need for them to undergo numerous invasive brain operations.

Capadona has discovered that it's more than just scarring that causes array failure. "We have

**"It's not only functional movement, but everything else that comes with that. It's very important to get them back on their feet."**

neurons dying around the implant," he says, and "we see that the materials of the implants are corroding and falling apart". He has traced these effects to a common source: the reactive oxygen molecules that are released as part of the body's inflammatory response. Capadona is now looking for drugs that might reduce that response, as well as developing coatings for the arrays that would act as an antioxidant, converting the oxygen species into water. The Department of Veterans Affairs is funding a preclinical trial of these coatings on the Utah array.

Capadona says that, ideally, he would like to build an entirely different implant from a polymer that is stiff enough to be manipulated easily during surgical insertion, but then softens as it absorbs water from the brain — therefore placing less mechanical stress on the tissue. It could also be laced with a drug that suppresses the initial immune response.

Another type of implant has been proposed by Charles Lieber, a chemical biologist at Harvard University in Cambridge, Massachusetts. In 2015, Lieber created a mesh of metal nanowires, which he then coated in a polymer. In solution, the mesh curls up into a cylinder that can be drawn up into a hollow needle and injected directly into the brain, where it can unfurl. Because the mesh is flexible and there is plenty of space between the wires, it does not exert the mechanical stress that leads to tissue damage. "It's fundamentally different from normal probes in that it doesn't elicit this response," he says. "You're not putting a thorn in the brain any more."

Lieber has tested the mesh in mice, where it was used for a year without degradation. He plans to do some initial tests in people with temporal-lobe epilepsy this year. Because one treatment for the condition involves removing part of the brain, the implant will be tested by attaching it to tissue that will be removed an hour or so later, rendering any damage from the device inconsequential.

Plugging computers into the brain to reinstate the ability to move is still just research with promise, rather than a practical treatment. Ajiboye has implanted arrays in only a dozen people in as many years. But, already, it's clear that restoring movement in people with paralysis gives them back more than just control of their limbs. Kochevar noted that the work he was involved with had given him a more positive outlook, and Contreras-Vidal says he's seen a change in the people whom his exoskeletons have helped to walk again. "They feel better psychologically, they are at eye level, they have better bladder function, they have less infections, they have better bowel movements, they have less skin conditions, they gain strength," he says. "It's not only functional movement, but everything else that comes with that. It's very important to get them back on their feet." ■

Neil Savage is a freelance writer in Lowell, Massachusetts.





Weight-bearing exercise can help to stave off the age-related loss of skeletal muscle.

#### AGEING

# Lifting the burden of old age

*The loss of muscle and strength that accompanies ageing can be debilitating. But is the inevitable process actually a disease that could be treated?*

BY LIAM DREW

At the time of Queen Victoria's accession to the British throne in 1837, the longest life expectancy for women in the world's most developed countries was roughly 45 years. By 2015, it had increased to almost 87 — a gain of more than 2 years a decade.

Much of this improvement is the result of profoundly lower rates of child mortality. But something else has also changed: old people are living for longer. "Since 1950, there has been enormous progress in bringing down death rates for people in their sixties and seventies and eighties," says James Vaupel, who studies ageing and the structure of populations at the Max Planck Institute for Demographic Research in Rostock, Germany.

The size of the elderly population worldwide is unprecedented, and the oldest of this group

are the fastest growing segment of society. In 2000, 71 million people were over the age of 80, according to the United Nations Department of Economic and Social Affairs. By 2030, that number will have increased to almost 202 million people, and by 2050, to 434 million.

This demographic shift poses profound questions as to the ability of medicine to meet the health needs of the oldest strata of society. "The paradigm of medicine has been curing, so the main issue has been mortality," says Alfonso Cruz-Jentoft, a specialist in geriatric medicine at the University Hospital Ramón y Cajal in Madrid. He thinks that needs to change. As people age, he explains, "function becomes more important than mortality". In other words, maintaining the ability to live independently may trump the need to prolong life for the very elderly. "The most meaningful definition of health is can you take

care of yourself," says Vaupel.

Few conditions are more central to the erosion of elderly people's independence than sarcopenia — an age-related loss of skeletal muscle mass and function. Progressive loss of such muscle can prevent a person from leaving their home, climbing stairs or even rising from their chair. These failures in daily living, as well as the falls that are associated with muscle weakness, are among the leading causes of admission to nursing homes and hospitalization among the elderly.

Recognition of sarcopenia as a condition of considerable concern for public health is, however, a fairly recent development. "We physicians all know about renal insufficiency and heart failure and respiratory failure," says Cruz-Jentoft, "but we'd never thought about muscle failure." It was only in 2016, when sarcopenia was officially recognized by the

World Health Organization's International Classification of Diseases, that doctors could formally diagnose people with the condition.

Even in the light of these positive steps, sarcopenia remains a condition with neither an agreed-on definition nor an effective treatment. As the average age of the world's population increases, researchers are working on both. "We know we're an increasingly ageing population," says Elaine Dennison, an epidemiologist who works on sarcopenia at the University of Southampton, UK. "One of the challenges for us is how to make sure that those added years are quality years."

### IN ALL BUT NAME

Sarcopenia's emergence as a clinical concern can be traced to a specific event. In 1988, Irwin Rosenberg, the then-director of the Jean Mayer USDA Human Nutrition Research Center on Aging at Tufts University in Boston, Massachusetts, attended a scientific meeting on health in older people in Albuquerque, New Mexico, after which he was asked to write up his notes<sup>1</sup>. In these, Rosenberg called attention to a point that clearly had been neglected, given that it touched so many aspects of health. "No decline with age is more dramatic or potentially more functionally significant than the decline in lean body mass," he wrote. "Why have we not given it more attention?"

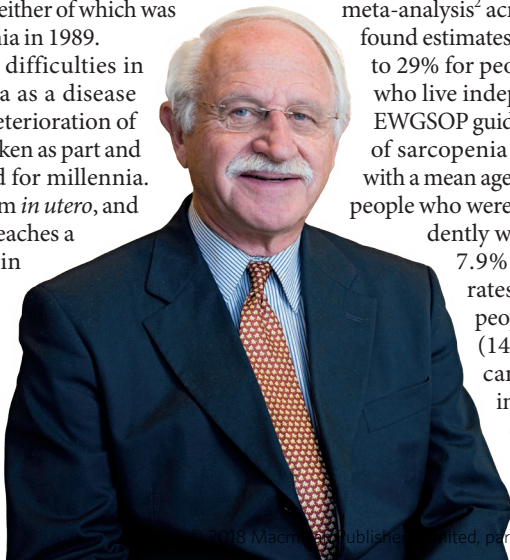
In answer to himself, and somewhat tongue-in-cheek, he offered: "Perhaps it needs a name derived from the Greek. I'll suggest a couple: sarcomalacia or sarcopenia."

Although the term sarcomalacia sank without trace, within a year, sarcopenia — meaning a loss or poverty of flesh — was the subject of a call for grant proposals by the US National Institutes of Health. "It was a pick-up of almost dizzying speed," Rosenberg says.

If the 27 years between the coinage of the term sarcopenia and recognition of the condition by the World Health Organization feels less dizzying, it is probably because establishing a disease category takes time. Before the medical community can develop treatments and prevention strategies, robust diagnostic criteria and the underlying disease-causing processes must be defined — neither of which was in place for sarcopenia in 1989.

One of the main difficulties in defining sarcopenia as a disease is that a degree of deterioration of the body has been taken as part and parcel of getting old for millennia. Muscle begins to form *in utero*, and then grows until it reaches a peak mass, usually in a person's late 20s. From then on, there is continual loss.

Irwin Rosenberg  
coined the term  
sarcopenia.



Although slow at first, with age the process quickens until, in some people, it reaches a level that impinges on daily life.

The stereotypical profile of the gain and subsequent reduction of muscle mass across a person's life echoes the life course of many tissues, and it presents a challenge to sarcopenia researchers: if all people can expect some natural loss, how severe must the loss be before it is considered a disease?

The European Working Group on Sarcopenia in Older People (EWGSOP) — a consortium, led by Cruz-Jentoft, of representatives from four major European science bodies working on ageing — was set up in 2009 to precisely define sarcopenia, facilitating basic and clinical research into the disease. EWGSOP published its initial guidelines for describing and diagnosing the condition in 2010, and similar groups in the United States (the International Working Group on Sarcopenia in 2011) and Asia (the Asian Working Group for Sarcopenia in 2014) have also produced recommendations. The goal of these collaborations was to come up with quantifiable metrics that would enable doctors to "decide who has sarcopenia and who does not," says Roger Fielding, a physiologist and colleague of Rosenberg at Tufts, who co-led the US effort.

The working groups agreed that sarcopenia should be defined not solely by muscle loss, but also by a measure of muscular function. To that end, they all recommended that an assessment of grip strength and gait speed should be part of the diagnostic procedure. However, when attempting to define cut-off points for speed and strength, below which a person can be said to have the condition, the three groups diverged — not drastically, but enough to prevent the adoption of a standard definition (see 'Crossing the threshold').

This has been problematic, says Dennison. "You need the definition to be able to do good studies, to look at the extent of the problem. And regarding treatment, you have to have hard end points to trials," she says. Estimates of the prevalence of sarcopenia have varied considerably, depending on both the definition

used and the population surveyed — a 2014 meta-analysis<sup>2</sup> across several countries found estimates that ranged from 1% to 29% for people aged 60 or older who live independently. Using the EWGSOP guidelines, the prevalence of sarcopenia in a UK population with a mean age of 67 and comprising people who were able to live independently was 4.6% for men and 7.9% for women<sup>3</sup>. Such rates are much higher in people in residential care<sup>2</sup> (14–33%), in those with cancer<sup>3</sup> (15–50%) and in patients in intensive-care units<sup>4</sup> (60–70%).

When developing

diagnostic parameters, many specialists in sarcopenia draw analogies with the recognition of osteoporosis as a disease in the 1980s. Similarly to muscle mass, bone density decreases with age from a peak value attained in a person's 20s, and tends to decline steeply in women after the menopause. However, to robustly demarcate osteoporosis as a medical condition, a cut-off needed to be set. This was done by plotting bone density against the risk of fracture, which rises as the density falls — slowly at first,

**"It may be that muscle is setting the pace of ageing of other tissues."**

but then increasingly dramatically. A density value at which the fracture risk was viewed to be unacceptably high was then picked. "It isn't that something magical happens when you hit

that threshold," says Dennison. But the threshold is tied to real-life outcomes — in the same way that the blood-pressure values used to define hypertension are linked to an elevated rate of adverse cardiovascular events. In both cases, crossing the threshold is a cue for medical intervention.

The quest to find a concrete link between muscle decline and real-life outcomes took a considerable step forward in 2012, according to Fielding, when epidemiologists involved in the US Foundation for the National Institutes of Health Sarcopenia Project presented a review of medical data gathered from more than 26,000 elderly people. They had set out to determine which clinically measurable parameters — be it degree of muscle loss or decline in physical performance — were most strongly linked to real-life outcomes such as slow walking or being unable to rise from a chair unaided. Such analyses are feeding into ongoing attempts to develop an internationally accepted definition of sarcopenia, and further guidelines are expected, including a revision from EWGSOP in late 2018.

However, agreeing on a disease threshold is unlikely to end the question of how to define muscle health in ageing. "Usually with new diseases, you start with the sickest patients," says Cruz-Jentoft, "before you move to intermediate ones." As the field of sarcopenia evolves, the threshold could fall or an 'at-risk' category could emerge — similar revisions have occurred, for example, for both hypertension and diabetes. Such progression will be shaped by a greater understanding of the underlying biology and the risk factors of sarcopenia, as well as — most importantly — the development of effective treatments. "A fundamental requirement of screening or early identification of a disease process," Rosenberg says, "is that you have something to offer."

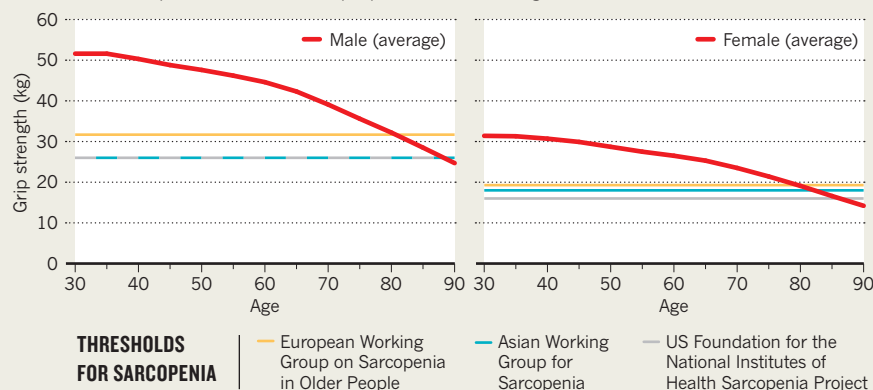
### HALTING SARCOPENIA

Sarcopenia does not have an unambiguous biological hallmark. There is no single process that is responsible for the demise of muscle



## CROSSING THE THRESHOLD

A person's grip strength begins to decline at around the age of 30, with an acceleration in his or her 60s. Working groups established to define sarcopenia do not agree fully on the point at which low strength should be considered a feature of the disease. But by the age of 85, most people's strength will fall below a clinical threshold for sarcopenia. However, not all people with such a strength will meet the full criteria for the disease.



fibres with age. Factors that contribute to the development of sarcopenia include hormonal changes (in particular, falling levels of testosterone, oestrogen or growth hormone), loss of the neurons that stimulate the muscle, an infiltration of fat into muscle, insulin resistance, physical inactivity, a vitamin D deficiency and not eating enough protein. And it's probable that the relative contribution from each varies between individuals.

Researchers who hope to prevent sarcopenia are looking for areas in which changes in lifestyle can make a difference. Trials investigating the use of weight or resistance training have yielded positive results, but aerobic activities alone have little impact. This has been demonstrated by numerous small- and medium-sized clinical trials, and resistance training is now being investigated further in combination with nutritional intervention in a large multicentre European trial.

Diet is another prominent modifiable factor. In particular, accumulating evidence indicates that eating too little protein can contribute to muscle loss. In 2013, a review led by the European Geriatric Medicine Society suggested an increase in the recommended amount of protein that people aged over 65 should consume, and advocated that older people who were ill should consume more protein still.

One group at particular risk of developing sarcopenia is older people who undergo long periods of inactivity as a result of, for example, serious illness or the need for sustained bed rest. Fielding advocates for making muscle rehabilitation an intrinsic part of managing the recovery from such an episode.

Other processes that lead to muscle loss seem to be intrinsic consequences of ageing, which require pharmacological intervention. Small biotechnology firms and large pharmaceutical companies alike have been active in this area of research for a decade, developing an array of compounds that act through

various mechanisms. Drugs that increase the sensitivity of the androgen receptor for the hormone testosterone have shown promise in phase II clinical trials. (However, simply administering testosterone to boost muscle mass causes a number of adverse side effects.) Researchers are also targeting a molecule called myostatin — one of hundreds of signalling molecules, known as myokines, that are released from muscle. Feedback from myostatin inhibits muscle growth, and drugs that block it have produced promising results in phase II trials.

Yet drug development is still at an early stage. These compounds bid to increase muscle mass and to stimulate muscle growth, but it's unclear whether this is the best approach to improving muscle function in ageing bodies. "The right target and the right mechanism of action is probably still unknown," says Fielding. He emphasizes that research into the processes underlying muscle decline in extreme old age remains immature because, until recently, people used to die earlier in life from other diseases, meaning such processes "weren't even in the wheelhouse of things to investigate".

One idea in its infancy is that treating sarcopenia could have broad anti-ageing effects. After myokines are released from muscle, they enter the bloodstream and regulate the activity of many organ systems. Fabio Demontis, who studies ageing at St Jude Children's Research Hospital in Memphis, Tennessee, is investigating whether changes in the levels of myokines released, owing to muscle ageing and inactivity, can affect the health of other tissues.

Demontis conducts his work in the fruit fly *Drosophila*, for which an arsenal of genetic tools enables researchers to perform experiments that are impossible in mammals at present. He is able to flick a genetic switch selectively in the muscles of these insects to slow muscle-fibre ageing — and has found that this action slows ageing in other tissues of the fly as well. "It may be," he says,

"that muscle is setting the pace of ageing of other tissues — and potentially of the whole organism."

Demontis is trying to find out whether the widespread effects of muscle ageing that he sees in flies are found in species throughout the animal kingdom. "Anything that helps with these devastating age-related conditions is very important," he says, echoing the views of clinicians. "If you are able to delay disease onset for ten years, that's a big deal."

## HOW TO TREAT AGEING

Rosenberg attributes the interest in muscle ageing that followed him naming sarcopenia to one main thing: "We take disease seriously, whereas we view processes of ageing as simply being natural."

David Gems, who studies the biology of ageing at University College London, thinks that there is nothing benign about senescence. He sees the myriad changes that occur throughout the body after the age of about 30, and that accelerate with age, as precursors to the outright diseases of old age. "I don't see how the idea that they're somehow non-pathological can stand up to rational analysis," he says. Convention, he argues, plays a large part in shaping what is viewed as normal in medicine.

But such views are fluid, and in the emergence of sarcopenia, both Gems and Rosenberg see a parallel with Alzheimer's disease. Gems says that when he was growing up in the 1970s, dementia was seen as a normal cognitive decline that came with age. "It was seen as nature taking its course," he says. "Granny's just having a second childhood, she's a little bit gaga."

But around the same time, neurologists revisited German psychiatrist Alois Alzheimer's work from the early twentieth century and reconceptualized dementia. Suddenly, stark cognitive decline — regardless of when it started — came to be viewed as a disease that might be halted. Rosenberg says that this led to "a meteoric rise, not only in interest, but in research funding and diagnosis". What constituted normal ageing for one generation had been redefined as an illness for the next.

When Rosenberg coined the term sarcopenia, he described it as "an opportunity". Although dementia has remained the most stubborn of foes, it's hoped that by focusing research on muscle ageing, the quality of life of people in their later years can be improved considerably. ■

**Liam Drew** is a freelance writer based in London.

1. Rosenberg, I. H. *Am. J. Clin. Nutr.* **50**, 1231–1233 (1989).
2. Cruz-Jentoft, A. J. et al. *Age Ageing* **43**, 748–759 (2014).
3. Patel, H. P. et al. *Age Ageing* **42**, 378–384 (2013).
4. Peterson, S. J. & Braunschweig, C. A. *Nutr. Clin. Pract.* **31**, 40–48 (2016).



a normal, dirty home environment". By the age of two or three, the composition of a child's gut microbiota is very similar to that of an adult's.

Should the assembly process be derailed, the consequences can be deadly. A considerably altered microbiota has been linked to a form of gut inflammation that is a leading cause of death in infants who are born prematurely. Less extreme changes to the microbiota in otherwise healthy babies might have long-term consequences for health, perhaps playing a part in conditions such as asthma and diabetes.

Researchers are looking for ways to rebalance the microbiota in premature infants. And some are wondering whether it might be possible to reshape the microbial community of the healthy infant gut to help prevent chronic diseases in adulthood.

### EARLY PERILS

Premature infants are especially vulnerable to disruption of the microbiota. Many are delivered by caesarean section, and therefore do not come into contact with the microbes that live in the birth canal. Such babies are also often given courses of powerful antibiotics and housed in sterile plastic incubators where they have minimal contact with human skin. Given that these interventions separate babies from their environment, it's not surprising that the gut microbiota of premature infants is markedly different from that of babies born at full term. It tends to have a lower proportion of microbes that are beneficial to gut health, such as *Bifidobacterium* and *Lactobacillus*, as well as a greater abundance of disease-causing bacteria and a lower diversity of bacteria in general. And the bacterial community is often chaotic, with dramatic shifts in composition over a matter of days.

The abnormal gut microbiota of premature infants is thought to have a role in their vulnerability to necrotizing enterocolitis, a severe form of gut inflammation that strikes suddenly in the first few weeks of life and can cause permanent damage to the intestine. Although full-term babies can develop the condition, at least three-quarters of cases occur in infants born prematurely. In the past two decades, as doctors have learnt to manage the respiratory problems of premature infants more effectively, necrotizing enterocolitis has become a main threat to such babies.

The cause of necrotizing enterocolitis isn't a particular microbe, but rather a dysfunction of the gut microbiota as a whole. As well as its role in digestion, the gut is an immune organ, says Barbara Warner, a neonatologist at Washington University in St. Louis. Early interactions of the gut with microbes are therefore powerful shapers of a child's immune system.

Necrotizing enterocolitis could be the consequence of this process going awry — perhaps representing "the baby's immune system struggling to work out what's the right thing to do", Embleton says. "Probably, this disease we see

Ingredients in breast milk can help to establish a healthy community of microorganisms in the infant gut.

### MICROBIOTA

# Baby thrivers

*Is a person's future health shaped by microorganisms encountered early in life?*

BY SARAH DEWEERDT

Within a few weeks of being born, a baby is host to a community of billions of bacteria, viruses and fungi — most of which are found in the gut — that can shape many aspects of health. How that community, or microbiota, assembles is a matter of debate: some researchers have begun to question the dogma that the womb is a sterile environment. Yet it's clear that birth sets off a radical transformation of the infant gut.

"It's an incredible ecological event," says

Phillip Tarr, a paediatric gastroenterologist at Washington University in St. Louis, Missouri. Colonization of the gut begins in earnest when a baby encounters microorganisms from its mother's vagina during birth. As the baby suckles at the breast, it picks up more microbes from its mother's skin. It also consumes microbes from its mother's gut that have infiltrated her breast milk.

Later, microbes are picked up from adoring visitors or a lick from the family dog, as well as what Nicholas Embleton, a neonatologist at Newcastle University, UK, refers to as "living in



is a sort of exaggerated inflammatory condition challenging a very immature and naive gut immune system.”

Treatment for the condition “is very, very crude and basic,” Embleton adds. Some babies with necrotizing enterocolitis can be treated with antibiotics and a temporary switch to intravenous feeding to give the intestine time to heal. More-severe cases require surgery to remove the damaged portion of intestine. The loss of a large part of the intestine can lead to lifelong difficulties with feeding or absorbing nutrients. About one-quarter of babies who develop the condition will die.

But now, researchers are looking to the gut microbiota for ways to stop the condition taking hold. Some are searching for clues that could help to predict the development of necrotizing enterocolitis, enabling earlier medical intervention. For example, an overgrowth of bacteria from the phylum Proteobacteria can precede the condition. But these microbes are also found in healthy infants, so it's not always clear when to sound the alarm. And such changes in microbiota composition might not be the true cause of the illness.

## FORTIFIED MILK

Breast milk might hold a solution. Since the 1990s, several studies have shown that breastfed babies are less vulnerable to necrotizing enterocolitis than are those fed with formula milk. A subsequent flurry of research into the relationship between breast milk and gut microbes found that breast milk contains ingredients that promote the establishment of a healthy gut microbiota.

One example is short chains of sugar molecules known as human-milk oligosaccharides. “They’re the second-most-abundant carbohydrate source in human milk after lactose, but they’re not for nutrition of the babies,” says Victoria Niklas, a neonatologist at the University of California, Los Angeles. Instead, these oligosaccharides provide food for helpful microbes such as *Bifidobacterium*. They also coat the lining of the gut and bind to pathogenic bacteria, making it more difficult for disease-causing microbes to invade.

Another component of breast milk, the protein lactoferrin, has a number of antimicrobial properties. It suppresses the growth of bacteria and can even trigger the death of certain harmful microbes by binding to inflammatory molecules called lipopolysaccharides.

Offering support to mothers of premature infants who wish to breastfeed might therefore help to promote a healthy gut microbiota and prevent necrotizing enterocolitis. A further potential strategy is to supplement the diets of early babies with human-milk oligosaccharides or lactoferrin. Several trials of such supplements have been completed and more are under way. Biotechnology companies are also developing supplements that contain key components of breast milk. (Niklas is chief medical and

scientific officer of one such company, Prolacta Bioscience in Duarte, California.)

Another approach to fighting necrotizing enterocolitis is to feed beneficial bacteria, or probiotics, to premature infants. The goal “is to try and mimic what happens in healthy, full-term, breastfed babies,” says neonatologist and researcher Keith Barrington at the University of Montreal in Canada.

In 2011, the neonatal intensive-care unit at Sainte-Justine University Hospital Center, where Barrington works, began to routinely feed probiotics to babies born before 32 weeks’ gestation. The infants received a cocktail of four species of *Bifidobacterium* and one of *Lactobacillus*, and the incidence of necrotizing enterocolitis fell by around 50%. More than half of the neonatal intensive-care units in Canada have followed suit in providing probiotics, with similar results. However, it’s not a perfect solution. Barrington’s team has shown that the probiotic strains are present in stools of the premature babies, which indicates that the microbes are able to grow in the infant gut. But these babies still have fewer beneficial bacteria and more pathogenic bacteria in their gut than do healthy, full-term breastfed babies. Combining probiotics with molecules such as human-milk oligosaccharides or lactoferrin might help to improve the picture, Barrington says. He plans to compare the effects on the gut microbiota of the combination of probiotics and lactoferrin with those of the probiotic treatment alone.

The neonatology community is divided on the role of probiotics in preventing necrotizing enterocolitis. “Half of us think that they’re probably a good idea and half think that the case isn’t proven yet,” says Embleton. “And even if we were to use probiotics, we really don’t know which ones we should be using and how much we give,” he says.

## MICROBIAL IMPACT

As the debate continues, researchers are investigating whether having the correct gut microbes might also be crucial to enabling healthy infants to thrive. For example, children delivered by caesarean section have a different gut microbiota from those born vaginally. Breastfed and formula-fed babies also have distinct microbiotas in their gut. Epidemiological studies suggest that caesarean delivery and formula feeding are associated with an increased risk of obesity and asthma, as well as other conditions, and many researchers think that these effects might be shaped by the gut microbiota. Could the infant gut microbiota therefore hold the key to preventing such conditions in later life?

The links are not straightforward. “These are complex problems and I think, to be honest, the microbiota is just one piece of it,” says

Warner. However, she adds, the microbiota is an attractive target for intervention because it might be easier to modify than other risk factors for certain conditions.

Some doctors have advocated, for example, that babies born by caesarean section be swabbed with a sample of their mother’s vaginal microbiota. But if that microbiota helps to promote a condition such as obesity, the intervention could have a downside. And if the mother harbours disease-causing bacteria, it could even be dangerous.

Few studies have been able to demonstrate the ability of probiotics to make a lasting change to the infant gut microbiota. “It’s extremely difficult to engineer microbial populations that will stick and benefit the host,” says Tarr. When the probiotics are discontinued, the gut microbiota usually reverts to its previous state with a matter of days.

But there could be progress on that front. In a 2017 study (S. A. Frese *et al.* *mSphere* 2, e00501-17; 2017), researchers from the University of California, Davis, and biotechnology company Evolve BioSystems of Davis, California, reported that breastfed infants who were given strain EVC001 of *Bifidobacterium longum infantis* still had the microbes in their guts 30 days after treatment with the probiotic had been stopped. This strain, which was developed as a probiotic supplement to breastmilk for babies by Evolve BioSystems, is extremely efficient at consuming human-milk oligosaccharides, says neonatologist Mark Underwood, who led the study. (Underwood has no financial interest in the company.)

“We thought, maybe we can make a big difference in this [microbial] community by — instead of keeping them on probiotics forever — treating them for a short period of time with probiotics, but then giving these beneficial bacteria a food source that they are uniquely capable of consuming,” Underwood says.

The babies seeded with *B. infantis* also had fewer pathogenic bacteria and more beneficial metabolites in their gut than did breastfed babies who did not receive the probiotic. This suggests that the microbiotas of healthy breastfed infants, used as a benchmark for studies in premature babies, are also ripe for improvement.

How far such improvement could go is uncertain. The *B. infantis* study is only a first step, and researchers are unsure about what an ideal neonatal gut microbiota would look like. Yet the growing importance of the microbiota is changing the approach of the doctors who care for the youngest patients. Among the medical specialities, “neonatology has never been at the top of the food chain,” Niklas says. “But it has now become abundantly clear that our practices and our interventions really hold the seed of future health.” ■

**Sarah DeWeerd** is a freelance science writer in Seattle, Washington.



Residents of Flint, Michigan, march in 2016 to demand that lead water pipes in the city be replaced.

#### PREVENTIVE MEDICINE

# Cleaning up our future health

*Can an evidence-based approach help to strengthen the case for limiting people's exposure to toxic chemicals?*

BY KARL GRUBER

Pollutants are everywhere. They can be found in the water that we drink, the air that we breathe and the food that we eat, and they are taking a toll on our health. In 2015, pollution was estimated to have caused almost 9 million deaths worldwide — three times more than those from AIDS, tuberculosis and malaria combined.

Pollution can have a negative impact on health at any point in a person's life. Often, the full effects are not seen for decades. Unborn babies and young children, for example, are especially vulnerable to the effects of methylmercury, a widespread pollutant that accumulates in fish and seafood and can cause intellectual disability and vision and hearing losses. According to a 2013 study (T. M. Attina & L. Trasande, *Environ. Health Perspect.* **121**, 1097–1102; 2013), exposure to lead in childhood had a negative effect on IQ that resulted in an economic cost to low- and middle-income countries of around 977 billion international dollars (a unit of currency devised to account for differences in purchasing power between countries). And in the past two decades, evidence that exposure to particulates in the air are linked to dementia has begun to build.

Mercury, lead and air pollution are found throughout the environment. They are among ten pollutants highlighted by the World Health Organization as chemicals that pose a considerable threat to public health. The neurological problems that they can cause, for which treatment is often lacking, are especially concerning. "In the past decade, there has been a steady increase in the incidence of neurological disorders, and a great deal of these brain problems have been linked to exposure to different pollutants," says Philip Landrigan, a paediatrician and epidemiologist in the Icahn School of Medicine at Mount Sinai in New York.

Although the risks they pose are great, there is little understanding of the effects on health of many common chemicals. Since 1950, more than 140,000 new chemicals have been synthesized, of which around 5,000 are now ubiquitous in the environment. Despite people's regular exposure to these compounds, a wide-ranging study led by Landrigan reported that fewer than half of these chemicals have been tested for safety or toxicity in humans.

"The failure to test widely used chemicals for their potential toxicity represents a failure of governments to act on behalf of their citizens, and failure of the chemical-manufacturing industry to take responsibility for the products it produces," he says. "We are conducting a massive toxicological experiment in the world today and our children, our grandchildren and future generations are the unwitting, unconsenting subjects."

#### THE QUEST FOR EVIDENCE

Before the health burden of pollution can be reduced, the compounds responsible must be identified. Researchers gather such evidence



from two main sources. One is epidemiological studies that match exposure to a chemical — determined by its presence in the blood or urine — to the likelihood of developing a medical condition. The other is laboratory-based studies of a chemical's effects in animals. Together, data from these sources represent the bulk of the evidence that is used to build a case against a pollutant, and to convince policy-makers of the need to ban or restrict it. But the process takes time.

“It takes over a decade for adequate toxicological and epidemiological data to be amassed to even begin making rational decisions about a chemical's risk to human health,” says Jonathan Martin, a toxicologist at Stockholm University. In part, this is down to the interpretation of results. “Toxicological data can always be criticized because it is done in animals or cells with questionable relevance to humans,” Martin says. And even when an epidemiological study shows statistical associations between chemical exposure and adverse health effects, it cannot provide unequivocal evidence for causation on its own.

As a result, vast amounts of data must be collected to build a solid case for removing a chemical from the environment, including findings made in a variety of species of animal. “Only when the toxicological effects that are observed in animals are the same ones that show up in humans in many large and well-constructed epidemiological surveys is there enough information to perhaps take regulatory action against a chemical,” says Martin. The building of evidence is therefore just the start of the journey down the long road towards a chemical's withdrawal from use.

### SAFE EXPOSURE

Removing a pollutant from use entirely is difficult. Inaction by regulatory bodies is one issue that hinders the process. Lead — for many years, a common component of paint, water pipes and petrol — is now known to be highly toxic. In the past decade, thousands of studies have drawn links between lead exposure and the development of numerous health problems, including reduced cognitive function in children and adults. But although lead has now been banned in certain applications, evidence of its negative effects on health existed for many decades before policies on exposure were changed, even as safer alternatives were developed. “This long delay was the direct consequence of fierce opposition and incessant political lobbying by the lead industry,” says Landrigan.

A similar story lies behind the decision in March 2016 by the US Environmental Protection Agency (EPA) to allow farmers to continue to use the pesticide chlorpyrifos, in direct contravention of advice from the agency's own scientists. An agricultural ban on the pesticide — which was phased out of residential use in the United States in 2000, owing to its neurological effects — was opposed by manufacturers



**A hand-held X-ray fluorescence spectrometer is used to test for toxic chemicals in a child's boot.**

such as Dow AgroSciences of Indianapolis, Indiana, and industry groups such as the American Farm Bureau Federation, based in Washington DC. In the weeks preceding the verdict, Scott Pruitt, administrator of the EPA, is reported to have told the bureau that US President Donald Trump's administration was “looking forward to working closely with the agricultural community”. “The decision flies in the face of clear evidence,” says Landrigan. “As a paediatrician,” he adds, “I find Pruitt's decision to be scientifically reprehensible and morally repugnant.”

Even when regulatory bodies decide to take action against a pollutant, actually doing so often proves difficult. One of the trickiest aspects for researchers and authorities to negotiate is the level of exposure that is deemed ‘safe’.

The neurological effects of lead have been shown to occur even at what are considered to be low levels of exposure. “Decades of research demonstrate, quite conclusively, that there is no safe level of lead exposure,” says David Bellinger, a neurologist at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts. The result is that the greatest burden of disease associated with lead comes not from people who experience high levels of exposure, but from those who encounter only low or moderate levels in the environment. Although harm to an individual is greater at higher levels of exposure, low-level exposure is a bigger problem for health-care systems. “You are more likely to get sick or die from heavy exposure to lead, for example, but your case will be just one of a

handful,” says Bruce Lanphear, an environmental health researcher at Simon Fraser University in Burnaby, Canada. “A lot more people will get sick from low-to-moderate-level exposure.”

Known as the ‘prevention paradox’, this concept also applies to a number of other pollutants with no known safe level of exposure, including airborne particulates and asbestos. And it poses a considerable problem to regulatory bodies — with no safe level, everyone is at risk, and health-care systems are not set up to tackle such a wide-ranging problem, says Lanphear. Preventive interventions at the population level “are difficult to implement in a health system dominated by medical care which is designed to treat disease”, he says.

Although efforts have been made to remove lead compounds from paints and petrol in most high-income countries, it remains a persistent contaminant of the plastic polyvinyl chloride, brass tap fittings, children's toys and even food. A 1983 report by the UK Royal Commission on Environmental Pollution found that lead was so widely dispersed in the environment, owing to its extensive use during the twentieth century, that “it is doubtful whether any part of the earth's surface or any form of life remains uncontaminated by anthropogenic lead”.

Faced with the impossibility of fully cleansing the environment of lead, regulatory bodies instead try to minimize people's exposure to it by setting legally acceptable levels. But often, such levels are not based on scientific evidence, and what is deemed allowable can vary considerably from place to place. For example, the Australian standard for brass pipe fittings permits a lead content of 4.5%, whereas the equivalent US limit is 0.25%. “The problem with lead lies in the regulatory systems that insist on allowing theoretical, but not empirically supported, safe levels,” says Mark Taylor, who studies environmental contamination at Macquarie University in Sydney, Australia.

Therefore, despite decades of research into its effects and attempts at regulation, lead continues to permeate the environment and to cause serious health problems — even in the most advanced countries in the world. A 2016 study found tap water contaminated with lead in more than half of 212 homes tested in the state of New South Wales, Australia. And in 2014, a crisis was seeded in Flint, Michigan, when the city's water supply was switched to the Flint River. The local authority failed to treat water from this new source with an anti-corrosion agent, which resulted in lead leaching from water pipes. The result was broad public exposure to unsafe levels of lead — almost 900 times the legal limit, in some cases — with devastating consequences for many unborn children. The Flint crisis prompted a number of investigations into water quality and lead poisoning in other parts of the United States. “The results indicated that childhood lead exposure in Flint is more the norm than the exception,” says Bellinger.

**“We are conducting a massive toxicological experiment in the world today.”**



A sign warns that the pesticide chlorpyrifos has been applied to an orange orchard in California.

Indeed, he adds, many areas of the United States have an even greater prevalence of children with high levels of lead in their blood.

### CHEMICAL WHACK-A-MOLE

In some cases, strong action founded in evidence can be taken against a chemical pollutant and still not yield the health improvements that were intended. Bisphenol A (BPA) is a substance found commonly in the liners of food tins, plastic water bottles and even the thermal paper on which shop receipts are printed. BPA's structure enables it to mimic or block the action of hormones. This allows the molecule to interfere with the function of the body's endocrine system — the complex network of glands, hormones and receptors that link the brain to reproduction and metabolism.

The detrimental impact of BPA is well established; evidence of its hormone-disrupting capabilities began to emerge in the mid-1930s. Numerous studies of its toxicity in people and animals eventually led several manufacturers to remove the chemical from their products. Unfortunately, the compounds that manufacturers now use instead of BPA are not much safer. "Replacement chemicals may be as bad, or even worse, than BPA," says Andrea Gore, a toxicologist at the University of Texas at Austin. "The chemical industry has switched to other members of the bisphenol family, but recent studies testing these bisphenols show that they are also endocrine disruptors," she says.

"Some have made the analogy to chemical 'whack-a-mole', whereby we try to increase human safety by regulating one chemical, only to see several similar chemicals pop up

to replace it," says Martin. Comparable issues have been reported for other toxic chemicals, including phthalates, per- and polyfluoroalkyl substances and flame retardants.

In each case, the initial response of manufacturers was to replace a regulated chemical with something similar. And because the properties of a material are linked directly to its chemical structure, such replacements often had similar effects on health. "It's easier and faster for the manufacturer to replace the restricted chemical with other existing chemicals having similar shapes, sizes and properties," says Martin. "This is not necessarily nefarious but it is dumb, and the best chemical regulatory systems in the world allow it to happen again and again."

### POLICY SHIFT

To minimize the adverse effects of pollution on health, many researchers recommend that people take matters into their own hands by limiting personal exposure to toxic chemicals. Exposure to BPA and its replacements can, for instance, be reduced considerably by avoiding tinned foods. And the consumption of pesticides can be reduced through careful preparation of fruit and vegetables, or by choosing organic produce grown largely without the use of synthetic substances.

But lifestyle changes have only a limited reach. "While behavioural modifications can reduce exposure, modifications in industrial practices are also likely to produce substantial reductions," says Leonardo Trasande, a paediatrician at New York University Langone Medical Center.

Some researchers are calling for the toxicity of a chemical to be established in advance of its introduction to the environment. "The idea that a chemical should be thoroughly tested for toxicity before it goes on the market seems like a no-brainer, but that simply isn't how we

regulate chemicals in the United States," says Gore. "The burden of proof should be on those who are profiting from the chemicals," she adds. Assessing the safety of a chemical takes a long time, however — an observation that led Philippe Grandjean, an environmental-health researcher at the University of Southern Denmark in Odense, to suggest lowering the bar researchers must clear to prove a chemical is unsafe. "We have to decide whether we need to wait many years or decades for solid proof, or if less documentation would be required in the interest of preserving the next generation's brain functions," he says.

To achieve considerable change in the way that pollutants are regulated worldwide, Gore thinks that pressure from the public will be crucial. "The way to change things is for people to vote for politicians who take strong pro-environmental stands," she says. Many of the positive steps already taken to protect people from harmful pollutants are rooted in heightened public awareness and the pressure that it exerts. BPA, for example, was banned from use in baby bottles by the US Food and Drug Administration (FDA) in 2012 after the American Chemistry Council requested the move to allay concern from the public. Between 2006 and 2015, per- and polyfluoroalkyl substances were phased out by chemical manufacturers and the EPA, against a backdrop of mounting public pressure that included a lawsuit against one manufacturer. And in Beijing, pressure exerted through Chinese social media — including the sharing of air-quality data collected by the US embassy — has been instrumental in prompting the authorities to begin to address the problem of air pollution.

"When science and court action builds enough pressure, it occludes industry voices," says Taylor. Dogged determination is required to ensure that promises are fulfilled, he adds, but the combination of incontrovertible evidence and public engagement offer the best hope of eliminating toxic pollutants and protecting health. "Humans are both the problem and the solution. The challenge facing scientists and policymakers is how to get the wider public to see and engage actively," he says.

In January, as a petition with more than 30,000 signatures circulated and demands from concerned customers mounted, a large chain of hardware stores in Australia and the United Kingdom revealed its intention to phase out the sale of a neonicotinoid insecticide that is linked to declining bee populations. "Now people are pressuring companies to ditch plastic, or taking legal action against governments to take new measures against long-overdue air-pollution problems," says Taylor. "Everybody can play a role in reducing pollution," he adds. "It is the cumulative impact of all our efforts that is important."

**Karl Gruber** is a freelance science writer in Perth, Australia.





Kiran Musunuru (centre) and his team are using genome editing in the mouse liver to modify enzymes that regulate levels of 'bad' cholesterol.

#### GENE EDITING

# The heart-disease vaccine

*Advances in gene editing raise the prospect of a one-off injection that could reduce the risk of cardiovascular disease.*

BY ANTHONY KING

Consider this scenario: it's 2037, and a middle-aged person can walk into a health centre to get a vaccination against cardiovascular disease. The injection targets cells in the liver, tweaking a gene that is involved in regulating cholesterol in the blood. The simple procedure trims cholesterol levels and dramatically reduces the person's risk of a heart attack.

According to World Health Organization statistics published in 2015, ischaemic heart disease and stroke are the leading causes of death worldwide. About 17.7 million people died from cardiovascular disease that year, and at least three-quarters of those deaths occurred in low- and middle-income countries. Although antibody-based therapies have been launched to help those most at risk, the cost and complexity of the treatments means that a simpler, one-off fix such as a vaccine would be of benefit to many more people around the world.

The good news is that a combination of gene

discovery and the blossoming of genome-editing technologies such as CRISPR-Cas9 has given this vision of a vaccine-led future for tackling heart disease a strong chance of becoming reality. The breakthrough came in 2003, when researchers investigated three French families with members who had potentially lethal levels of low-density lipoprotein (LDL) cholesterol and who harboured a mutation in the gene *PCSK9* (ref. 1). *PCSK9* encodes an enzyme that regulates levels of LDL — or 'bad' — cholesterol. The mutations uncovered in the families increased the enzyme's activity, raising the level of LDL cholesterol in the blood. Breaking *PCSK9*, so that the enzyme it encodes loses its function, might therefore reduce LDL-cholesterol levels.

Sensing the possibilities, investigators at the University of Texas Southwestern Medical Center in Dallas sought to determine whether naturally occurring mutations in *PCSK9* could also have the effect of lowering LDL cholesterol. The researchers interrogated the Dallas Heart Study, a landmark investigation of

cardiovascular health carried out from 2000–02 in 6,000 adults living in Dallas County. The participants recruited represent the three main ethnic groups of the United States. After combing the data from about 3,600 individuals who provided a blood sample, the researchers sequenced DNA from the 128 participants with the lowest levels of LDL cholesterol. They discovered that about 2% of African-American participants had one broken copy of *PCSK9*, resulting from one of two inherited mutations<sup>2</sup>. A follow-up study of a different, larger population similarly found mutations in almost 3% of African Americans, which was associated with an 88% reduction in the risk of ischaemic heart disease<sup>3</sup>. "I think of them as having won the genetic lottery," says Kiran Musunuru, who studies human genetic variation and the risk of heart disease at the University of Pennsylvania in Philadelphia.

Musunuru thinks that in the next 20 years, gene editing will enable researchers to confer a mutation in *PCSK9*, or other beneficial mutations, on people who have had less luck in the genetic sense. "They would be dramatically



protected against heart attack and stroke for the rest of their lives,” he enthuses.

Others are more bullish. Technologies for delivering gene editing can be safe, effective and work in the long term, says Sander van Deventer, operating partner at investment firm Forbion Capital Partners in Naarden, the Netherlands. van Deventer played an important part at uniQure in Amsterdam, where he supervised the development of alipogene tiparvovec (Glybera), the first gene therapy to gain regulatory approval. He thinks that gene therapy to reduce the risk of cardiovascular disease could become a reality within 5 years — initially targeted to help people with high cholesterol (a condition known as hypercholesterolaemia).

### THE GATEKEEPER ORGAN

The liver is a preferred target organ of gene therapy for companies such as Editas Medicine in Cambridge, Massachusetts, Sangamo Therapeutics in Richmond, California, and CRISPR Therapeutics, also in Cambridge; it is straightforward to deliver genes to the liver, and the CRISPR–Cas9 tool is especially efficient in the organ, editing a greater proportion of cells than it does in most other tissues. The liver is also an excellent place from which to tackle cholesterol — it clears LDL cholesterol from the blood and is also a main engine of lipid synthesis. “The liver is the gatekeeper for removal of excess cholesterol from the body,” says William Lagor, a molecular biologist at Baylor College of Medicine in Houston, Texas.

The enzyme produced by *PCSK9* causes receptors for LDL cholesterol, found on the surfaces of cells throughout the body, to move inside the cell. With fewer receptors available to bind such cholesterol, its level in the blood rises. Already, two antibody-based therapies have been developed to inhibit the enzyme *PCSK9*, increasing the number of LDL-cholesterol receptors and consequently reducing the amount of cholesterol in the blood. One such *PCSK9* inhibitor, evolocumab (Repatha), can cut the risk of heart attack by 27% and stroke by 21%, when administered in combination with statins. But the treatment involves regular infusions of drugs for the rest of a patient's life and costs about US\$14,500 per year, a price that many commentators have deemed too high.

In 2014, Musunuru and his team showed that more than half of *Pcsk9* genes in the mouse liver could be silenced with a single injection of an adenovirus containing a CRISPR–Cas9 system directed against *Pcsk9*. This led to a roughly 90% decrease in the level of *Pcsk9* in the blood and a 35–40% fall in blood LDL cholesterol<sup>4</sup>. Next, they used a mouse engineered to contain human liver cells, and tuned the CRISPR–Cas9 payload to target human *PCSK9* (ref. 5). The team succeeded in showing that the human gene can also be switched off. “I’m convinced that if we gave this therapy to a human, it would work,” Musunuru says.

The approach is “absolutely plausible, even

feasible”, from a technological point of view, says Lagor. But there is also a philosophical barrier to negotiate. “You don’t necessarily want to treat people who haven’t got a disease yet,” he says. Karel Moons, a clinical epidemiologist at University Medical Centre Utrecht in the Netherlands, goes further. “Changing lifestyle may be much more effective for a population than focusing on high-cost interventions,” he says. He worries that a gene therapy for individuals at high risk would hinder efforts to help people to help themselves. “It is the way the human mind works. Take a pill and we think we are protected,” he warns.

Musunuru accepts that the idea does not have universal approval but thinks that “there will be greater enthusiasm for human trials for common diseases after genome editing has been proven safe in the patients with grievous genetic disorders”. Debilitating single-gene conditions such as Duchenne muscular dystrophy are likely to be first to benefit from therapeutic gene editing (see ‘Benefits from a partial fix’). Musunuru suggests familial

hypercholesterolaemia — the LDL-cholesterol disorder characterized in the three French families — as a similarly logical place to start. The associated mutations in *PCSK9* raise LDL-cholesterol levels from birth, causing premature heart attacks — sometimes in childhood — in those who are worst affected. “It would make a lot of sense to knock out the faulty *PCSK9* gene in those patients,” he says.

People with hypercholesterolaemia can make changes to their lifestyle and diet, as well as take statins, but this is often not enough. They might also require treatment with antibodies directed against *PCSK9* and frequent cleaning of the blood to remove LDL particles. Those with the most severe disease would receive the greatest benefit from genome editing, says Musunuru, and be the first candidates for therapy. “The strongest rationale for using genome editing is that it would be given just once, whereas patients have to take antibodies every few weeks for the rest of their lives.” He views the approach as being particularly useful for people in low-income countries with less-well-funded

## DUCHENNE MUSCULAR DYSTROPHY

### Benefits from a partial fix

Duchenne muscular dystrophy is a single-gene disorder that will probably be in the vanguard of diseases targeted by gene therapy. The condition affects up to 1 in 3,500 boys and men, and causes the progressive weakening of muscles; heart-muscle failure is the leading cause of death in people with the disorder. “This disease has resisted every therapy applied to it,” says Eric Olson, a molecular biologist at the University of Texas Southwestern Medical Center in Dallas. “The only reasonable approach is to go to the root cause of the disease, to the mutated gene. CRISPR seems an ideal approach.”

At the core of the condition lie defects in dystrophin, a long membrane-associated protein that acts as a shock absorber in muscle cells (pictured). Dystrophin’s central portion comprises 20 or so repetitive sections, which are analogous to the coils of a spring. *DMD*, the gene that encodes dystrophin, is long, containing 79 coding sections, or exons, and Olson says that mutations anywhere along its length can eliminate the production of functional dystrophin.

Rather than correcting specific mutations, he estimates that 80% of patients could benefit from a partial fix. Some of the coils in dystrophin can be deleted without

destroying the protein’s function. This means that sections of DNA within *DMD* that contain mutations can be removed. The shortened gene will make a working, truncated protein. “One edit can bypass all the mutations,” Olson says.

Dystrophin production as low as 5% of the normal level is thought to improve muscle function; Olson thinks that reaching 15% would bring major clinical benefits. In 2017, researchers at the Ohio State University in Columbus blew past that target, restoring dystrophin-expression levels in the heart muscle of mice by up to 40%, simply by slicing out a defective portion of *Dmd* using a CRISPR–Cas9 system delivered by a viral vector<sup>14</sup>. “So long as the gene can still read out, you make a partially functional protein,” says Renzhi Han, who led the study. His lab is now evaluating the safety of the strategy in mice. Olson’s research group has used the technique to restore up to 90% of normal dystrophin levels<sup>15</sup>.

Han and others are optimistic that trials in people can begin in the next five years. “Duchenne is the most devastating muscle disease. There is no escaping the clinical consequences,” says Olson. “There is enormous excitement in the Duchenne community about this new technology.” **A.K.**



health-care systems: “I do not see daily pills or monthly injections as being a realistic approach in the developing world.” But although a one-off treatment should be cheaper, drug companies could be tempted to charge a high price, on the basis that it achieves the same effect as do decades of expensive antibody-based drugs.

For now, Musunuru says that we need to work out the safest way to perform gene editing in people — not necessarily CRISPR–Cas9 — and also the best way for it to be delivered. Regulatory approval for a clinical trial would then be required, which could take a few years to achieve.

### STACKING TARGETS

Since the discovery of *PCSK9*, other variants in genes that alter the risk of cardiovascular disease have emerged. Some affect triglycerides, the main component of fat in the body; high levels of triglycerides in the blood are a known risk factor for heart disease. Apolipoprotein C-III inhibits the breakdown of triglycerides by enzymes; a mutation in *APOC3*, the gene that encodes it, was discovered in a population of Amish people in the United States in 2008 (ref. 6). The 5% of the group who were carriers had lower levels of LDL cholesterol, higher levels of high-density lipoprotein (HDL) — or ‘good’ — cholesterol and lower levels of triglyceride in the blood, all of which might reduce the risk of cardiovascular disease. A similar pattern has also been found in people who carry the mutation in Crete, Greece.

Musunuru is optimistic that knocking out a gene called *ANGPTL3* can reduce levels of LDL cholesterol and triglycerides. He was part of a team that reported in 2010 on three generations of a family with mutations in *ANGPTL3* and that had no history of heart disease and had low levels of cholesterol and triglycerides in the blood<sup>7</sup>. In 2017, three family members who had a complete loss of function of the protein encoded by *ANGPTL3* were examined<sup>8</sup>. “As far as we can tell, they are substantially protected against cardiovascular disease, but suffer no harmful consequences whatsoever,” says Musunuru. At least 1 in 300 people has a broken copy of *ANGPTL3*, which has been shown to reduce the risk of ischaemic heart disease by roughly one-third<sup>9</sup>.

Another potential target is the gene *LPA*, which encodes lipoprotein (a). High levels of lipoprotein (a) are a main risk factor for heart disease and stroke, yet no treatments have been approved by regulators such as the US Food and Drug Administration specifically to lower its levels. “This really is an ideal candidate for disruption with a liver-directed CRISPR gene-editing approach,” says Lagor. Initial candidates for the treatment would be people with extremely high levels of lipoprotein (a) who also have cardiovascular disease.

The most effective treatments will probably disrupt several of these genes at once to provide the greatest benefit. “Since *PCSK9* and

*ANGPTL3* work by different mechanisms, in principle they should be additive,” says Musunuru. Lagor agrees, adding that there are also economic upsides. “It is likely that the cost of targeting two genes, or perhaps even three or four, would be the same as for one gene.”

### REASONABLE OPTIMISM

Before gene-editing therapy can become routine, two main safety concerns must be addressed. First, off-target effects can occur when the RNA molecule that guides the Cas9 cutting enzyme into position misidentifies its complementary sequence of DNA, resulting in cuts being made in the wrong place. Second, the cellular machinery that repairs the double-strand breaks created in the DNA during gene editing might make an unexpected deletion or addition. Such mishaps could lead to the development of cancer. And although a considerable degree of risk might be acceptable for seriously ill patients with no other option, preventive gene therapy must clear a higher bar. “If the vaccine is being envisioned for the general population, then it needs to be essentially 100% safe,” says Musunuru, “at least to the same degree as the infectious-disease vaccinations that are routinely given to infants and children.”

A new technology from chemical biologist David Liu’s laboratory at Harvard University in Cambridge, Massachusetts, has therefore excited those in the gene-editing field. Liu has developed a technique that uses a modified CRISPR–Cas9 system to alter individual pairs of bases in cells without having to break the DNA double strand<sup>10</sup>. His team was able to chemically change the DNA base cytosine (C) into uracil (a base found in RNA), which the cell later replaced with thymine (T). In 2017, Liu’s team created another tool that could rearrange an adenine (A) so that it resembled a guanine (G), and then hoodwinked the cell into fixing the complementary strand of DNA to make the edit permanent, therefore changing an A•T pair into a G•C (ref. 11).

“Base editing is as big a development as the original introduction of CRISPR–Cas9 to the genome-editing field,” says Musunuru. “It’s totally changed how I’ve been thinking about tackling cardiovascular disease — in a positive way.” He is planning to test Liu’s A-to-G base editor in mice to see how well it works.

Gene-editing researchers have embraced targeted base editing to install precise changes without the uncertainty that accompanies a double-strand break. The technique has been used in labs to correct genes in yeast, plants, zebrafish, mice and even human embryos. A proof-of-concept study by Alexandra Chadwick, a postdoctoral researcher in Musunuru’s lab, delivered a base editor into the livers of adult mice to disable *Pcsk9*, halving the level of *Pcsk9* and cutting LDL cholesterol by almost one-third<sup>12</sup>. Musunuru adds that he has preliminary results showing base editing of *Angptl3* in mice using Liu’s C-to-T method.

The pace of innovation in gene editing has created an aura of optimism, particularly around the treatment of people with genetic disorders who have few or no other options. “It makes sense to begin therapeutic efforts with such diseases, even if the understanding of all potential risks is imperfect,” says Liu. But there is the potential for the technique’s use in the clinic to spread beyond these testing grounds. van Deventer has successfully lowered LDL cholesterol in mice by silencing apolipoprotein B-100 using a method called RNA interference<sup>13</sup>; he sees great potential in using the microRNAs that underpin the technique and, eventually, gene editing to address heart disease. “ANGPTL3,

**“The strongest rationale for using genome editing is that it would be given just once.”**

*PCSK9* and *APOC3* are targets not easily addressed by small molecules or antibodies,” he says. And the one-off nature of gene-editing treatments cuts down on issues with patients not following advice about

when to take a drug — a perennial problem concerning people on long-term medication.

“If you are talking about cardiovascular disease as a global health threat, which it undoubtedly is, then protecting the entire population is what we need,” says Musunuru. Lifestyle changes are important, but a substantial portion of the risk of heart failure and stroke comes from the genome. “You don’t need to choose between medicine and lifestyle. You should be doing both,” says Liu, citing people with diabetes, who fare best when they take medication and adjust their lifestyle.

“To vaccinate large numbers of people, that is some way off,” says Musunuru. But gene editing could reset the odds for those who didn’t win the genetic lottery, he predicts. “One way or another, genome editing is going to underlie a host of new types of cardiovascular therapies over the next 25 years.” ■

**Anthony King is a freelance science writer based in Dublin.**

1. Abifadel, M. *et al.* *Nature Genet.* **34**, 154–156 (2003).
2. Cohen, J. *et al.* *Nature Genet.* **37**, 161–165 (2005).
3. Cohen, J. C., Boerwinkle, E., Mosley, T. H. Jr & Hobbs, H. H. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
4. Ding, Q. *et al.* *Circ. Res.* **115**, 488–492 (2014).
5. Wang, X. *et al.* *Arterioscler. Thromb. Vasc. Biol.* **36**, 783–786 (2016).
6. Pollin, T. I. *et al.* *Science* **322**, 1702–1705 (2008).
7. Musunuru, K. *et al.* *N. Engl. J. Med.* **363**, 2220–2227 (2010).
8. Stitzel, N. O. *et al.* *J. Am. Coll. Cardiol.* **16**, 2054–2063 (2017).
9. Koschinsky, M. & Boffa, M. *Expert Opin. Ther. Targets* **18**, 747–757 (2014).
10. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. *Nature* **533**, 420–424 (2016).
11. Gaudelli, N. M. *et al.* *Nature* **551**, 464–471 (2017).
12. Chadwick, A. C., Wang, X. & Musunuru, K. *Arterioscler. Thromb. Vasc. Biol.* **37**, 1741–1747 (2017).
13. Koornneef, A. *et al.* *Mol. Ther.* **19**, 731–740 (2011).
14. Refaey, M. *et al.* *Circ. Res.* **121**, 923–929 (2017).
15. Amoasii, L. *et al.* *Sci. Transl. Med.* **9**, eaan8081 (2017).



# Cancer's cost conundrum

*The price trajectory of oncology drugs is unsustainable — but fixes are in the works.*

BY ELIE DOLGIN

The year 2011 was a watershed for cancer medicines in the United States. In the space of five months, federal regulators approved the first checkpoint-inhibitor immunotherapy, the first treatment for an aggressive form of thyroid cancer, the first personalized drug for the skin cancer melanoma, the first in an innovative class of targeted agents for lung cancer, and a 'weaponized' antibody therapy that delivers a drug to tumour cells in people with lymphoma.

The potency, complexity and innovative nature of these treatments were noteworthy. But so was the price. Each cost more than US\$100,000 per person when taken for a year — a rarity at the time for oncology drugs.

The prices seemed staggering to doctors, patients and health-care providers alike. But quickly, they became normal. By 2014, the average cost of a new orally administered cancer medicine exceeded \$135,000 a year — up to six times the cost of similar drugs approved in the early 2000s, after adjusting for inflation<sup>1</sup>. 2017 brought the most eye-popping price tag in oncology yet: a one-time cost of \$475,000 per patient for a personalized cell-based therapy for childhood leukaemia.

This generation of treatment promises to transform the field of cancer, yielding more cures and long-term remissions than ever before. But as medicine's ability to tackle tumours races ahead, health-care systems worldwide are struggling to deliver the benefits. If the affordability of drugs is not addressed soon, many people with cancer might not be able to reap the rewards of cutting-edge therapies. "We're on a trajectory that's really unsustainable," says Ameet Sarpatwari, an epidemiologist and legal scholar who studies drug pricing at Brigham and Women's Hospital in Boston, Massachusetts.

"It's really a major issue," says Sabine Vogler, a health economist at the Austrian Public Health Institute in Vienna. Drugs are unaffordable in many parts of the world<sup>2</sup>. "We have to ask ourselves," she says, "how long can we continue paying these high prices?"

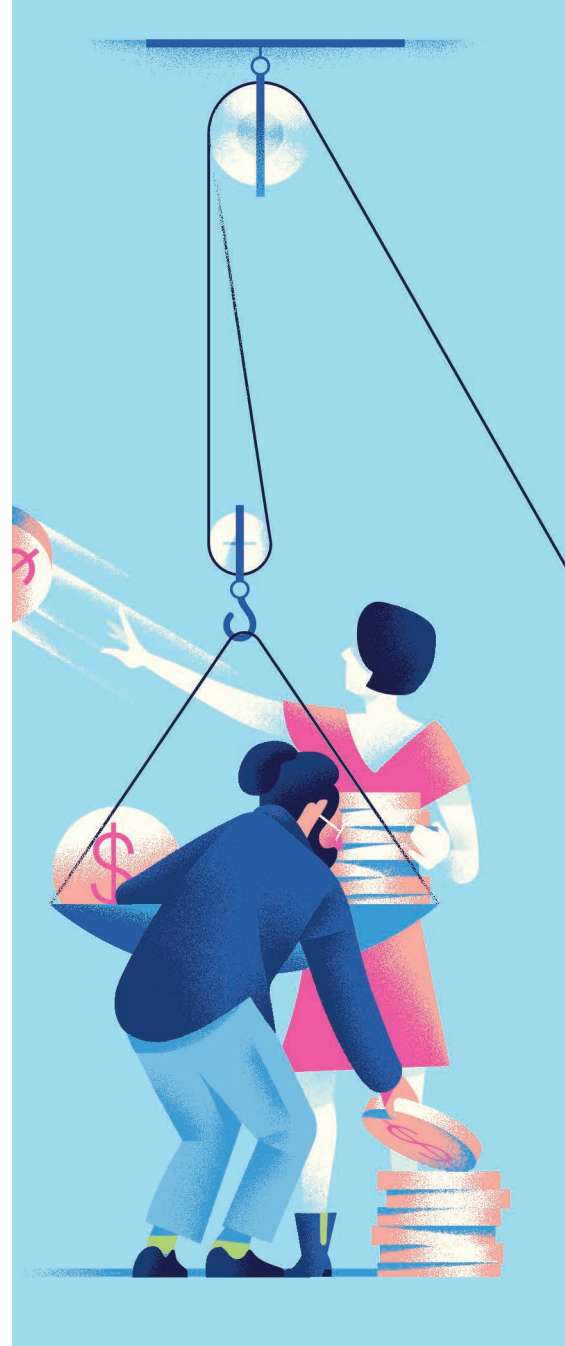
## STRATEGIES OF CONTAINMENT

New drugs are not the only aspect of cancer care that is getting more expensive. The costs associated with doctors' salaries, diagnostic tests, radiotherapy and surgery are all rising, says Darius Lakdawalla, a health economist at the University of Southern California in Los Angeles. Collectively, they continue to make up the lion's share of cancer-care expenditure. "This is a systemic problem," he says.

And as Daniel Goldstein, an oncologist and health economist at the Rabin Medical Center in Petah Tikva, Israel, and his colleagues reported last year, even the cost of existing cancer drugs has been increasing precipitously — well above the rate of inflation and much faster than other aspects of health care<sup>3</sup>. This price creep, as Goldstein calls it, can cause harm to patients, with a large number of them delaying or skipping treatments that they can no longer afford. Health-care costs are then compounded, Sarpatwari says, because people who don't take their drugs as scheduled are more likely to require hospitalization at a later point. "If people can afford their drugs, it can decrease downstream spending," he says.

The catalyst for spiralling costs starts in the United States, where the price of a drug "is not linked to anything rational", says Vinay Prasad, a cancer specialist at Oregon Health & Science University in Portland. This, he suggests, enables drug companies to charge exorbitant amounts for new treatments that are often not much better than older, cheaper options. And although other countries can usually negotiate a discount, the prices paid are often benchmarked against those in the United States. "What happens in America really has an impact on the rest of the world," Goldstein says.

One idea for lowering prices is to tie them to the level of clinical benefit provided. Peter Bach, a physician and cancer-drug pricing theorist at Memorial Sloan Kettering Cancer Center in New York City, has developed one such calculator of value-based prices: DrugAbacus. This online tool lets users calculate drug prices on the basis of their views on the relative importance of factors such as tolerability, new mechanisms of action, research and development costs and disease rarity — as well as the monetary value that



they place on a year of life. Under reasonable economic assumptions, DrugAbacus shows that 80–85% of cancer drugs are overpriced in the United States (see 'Over the odds').

Most countries with nationalized health care already have a value-based price-negotiation system in place — but even then, there are loopholes. In England, for example, the National Health Service spent almost £1.3 billion (US\$1.8 billion) between 2010 and 2016 on the Cancer Drugs Fund, a pot of money set aside to improve access to innovative treatments that ended up being used to pay for medicines that the country's drug-pricing watchdog, the National Institute for Health and Care Excellence (NICE), did not deem to be cost-effective.

An analysis<sup>4</sup> conducted in 2017 by Richard Sullivan, director of the Institute of Cancer Policy at King's College London, and his colleagues found that the fund had "not delivered meaningful value to patients or society". It has since stopped paying for drugs that were rejected by

MICHELE MARCONI





NICE, although it still covers medicines for which the institute's appraisal was inconclusive and further real-world data are required.

An alternative cost-cutting proposal takes the form of a money-back guarantee. Under such an arrangement, only those who obtain medical benefit from a drug have to pay for it. This kind of success fee could eliminate wasteful spending on drugs that do not work for a lot of people, but it has yet to do so in practice.

The best data on this sort of scheme come from AIFA, the Italian Medicines Agency, which introduced performance-based reimbursement for 25 cancer drugs in 2006. Two independent analyses<sup>5,6</sup> of the scheme suggest that it introduced extra layers of administration for little financial benefit. But the money-back guarantees "are still in place, despite their poor performance", says Livio Garattini, a health economist at the Mario Negri Institute for Pharmacological Research in Bergamo, Italy.

Some pharmaceutical companies are also beginning to offer this kind of guarantee on a voluntary basis. Novartis of Basel, Switzerland, for example, has said that people who receive tisagenlecleucel (sold as Kymriah), the company's \$475,000 therapy for leukaemia (available only in the United States, at present), can get a full refund if they show no improvement in the first 30 days after treatment.

"We are proud to offer this outcomes-based approach for Kymriah, which is unprecedented in this disease area," says Eric Althoff, head of global media relations at Novartis. He cites independent analyses by NICE and the Institute for Clinical and Economic Review in Boston, as well as evidence from economists at Novartis, to show that the price, even without a discount, is cost-effective for health-care systems. "We recognize our responsibility to ensure patient access and the need for a holistic, evidence-driven approach which incorporates clinical outcomes, patient experience,

benefit to the health-care system and societal value," Althoff says.

Goldstein, however, brushes off the guarantee as a public-relations stunt, rather than a real cost-containment measure. He points out that

**"What happens in America really has an impact on the rest of the world."**

the treatment fails in about 20% of people in the first month, which makes the average cost per person treated, after refunds, about \$380,000. That's almost the same as a similarly

effective treatment from Gilead Sciences in Foster City, California, which was approved in the United States just weeks after Kymriah, and has a price tag of \$373,000 per patient but no money-back guarantee.

"It becomes mathematical gymnastics," Goldstein says, with the cost of the guarantee baked into the list price. "It's all basically a little bit of a trick."



A Nepalese man receives cancer drugs provided through an access programme.

## GOVERNMENT INTERVENTION

More-radical steps could be taken to force down drug prices, even in the United States, where health care is largely a private, decentralized affair. Under federal law, the US government has the right to procure generic versions of patented drugs in exchange for 'reasonable' royalties that compensate patent holders.

According to a 2017 analysis<sup>7</sup> by Hannah Brennan and her colleagues at Yale Law School in New Haven, Connecticut, the US Department of Defense relied on this to obtain antibiotics and other drugs at steep discounts throughout the 1960s and early 1970s. And the threat alone of such action has been enough to rein in excessive drug pricing; in the wake of the 2001 anthrax attacks, a drug company fended off federal intervention by halving the price of its anthrax medicine.

"It's time to reconsider how the government provides medications," says Brennan, now an associate at the consumer-rights law firm Hagens Berman Sobol Shapiro in Cambridge, Massachusetts. "If drug companies are going to continue making up list prices and completely untethering them to anything," she adds, "then this is an appropriate and proportionate response."

The governments of countries in the European Union might be able to negotiate with drug companies to set prices, but they tend to do so in isolation, "which weakens the purchasing power," says Vogler. To address the problem, some EU countries have banded together to create a united front against pharmaceutical companies. The Netherlands, Belgium, Austria and Luxembourg have formed one such union. Half a dozen Mediterranean countries hope to do the same.

But even when drug companies do offer large discounts, there are many places in which cancer medicines remain out of reach. In several parts of Africa, for example, Swiss pharmaceutical giant Roche has engaged with governments and patient groups to provide

its breast-cancer drug trastuzumab (Herceptin) at half the usual price. That markdown was enough for the government of Kenya to agree in 2016 to foot the other half of the bill, at least for a small group of people. The country's Ministry of Health last year committed around 20 million Kenyan shillings (US\$195,000) to the effort.

The cost was too high for the government of the much poorer nation Rwanda, however. A 50% concession is "still so beyond what they can afford," says Lawrence Shulman, director of the Center for Global Cancer Medicine at the University of Pennsylvania Abramson Cancer Center in Philadelphia, who works in the East African country.

In a bid to pressure pharmaceutical companies into making expensive medications available to all people in low- and middle-income countries — as happened with HIV drugs in the 2000s — Shulman and an international team of leading cancer researchers worked with the World Health Organization (WHO) in 2015 to expand its list of essential medicines<sup>8</sup>. That helped to prompt two large pharmaceutical companies — Pfizer of New York City and Cipla of Mumbai, India — to agree in June 2017 to offer 16 medicines, most of which are on the WHO list and some of which were advocated by Shulman's group, at rock-bottom prices for people in Rwanda, Kenya and four other low-income countries in Africa.

But the drugs were all staples of chemotherapy treatment that are available as generic versions. Neither trastuzumab nor any other branded medicine from the list was included in the deal. For the most part, the newest therapies continue to elude those in need in the developing world, where a diagnosis of cancer means a painful and distressing death for most people (see "The diagnosis differential").

One notable exception is imatinib (Glivec). Since 2001, and essentially in parallel with the drug's first approval for use in chronic myeloid leukaemia (CML),

**"We felt this obligation to try to make the drug reach as many people as needed."**

Novartis has made the treatment available — at no cost — to the poorest people of the developing world through the Glivec International Patient Assistance Program. "This is a drug

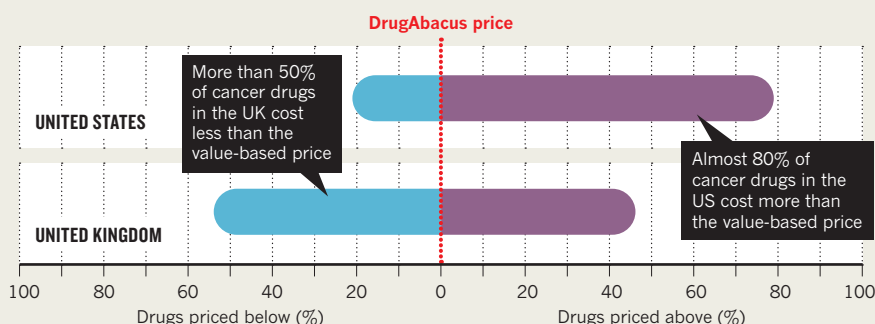
that gave people a normal life back," says David Epstein, a former chief executive at Novartis who is now at Flagship Pioneering, a venture-capital firm based in Cambridge, Massachusetts.

"We felt this obligation to try to make the drug reach as many people as needed," he says. The programme has handed out around 2.3 million monthly doses of the drug to more than 50,000 people in 80 countries<sup>9</sup>.

Building on that success, the Max

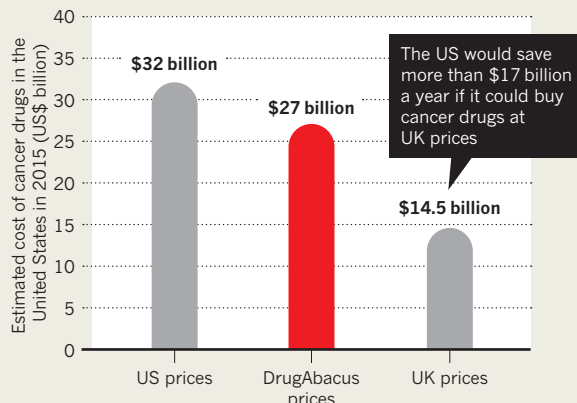
## OVER THE ODDS

Linking a drug's price to the clinical benefit that the medication provides — a practice known as value-based pricing — has the potential to reduce spending on cancer drugs. The DrugAbacus tool provides reasonable estimates of value-based prices\* and can be used to indicate whether cancer drugs are priced appropriately.



The total estimated spend on cancer drugs in the United States in 2015 was US\$32 billion — almost \$5 billion more than if the drugs had been purchased at the prices suggested by DrugAbacus. The same drugs would have cost only \$14.5 billion at UK prices.

\*The value-based pricing ('DrugAbacus price') used in the analysis assumes that an extra year of life is worth \$132,000 and that a 15% discount should be applied to drugs with severe side effects. Increasing the value of an extra year of life increases the percentage of drugs that are available at or below the DrugAbacus price. The data cover the prices of 52 cancer drugs in the US Medicare system and the UK National Health Service.





Foundation — a non-profit organization in Seattle, Washington, that runs the imatinib access programme in partnership with Novartis — has worked with three other manufacturers to make all five of the CML-targeted treatments on the market available in the same way. “Today, if you live in the lowest-income economies of the world and you’re diagnosed with CML, you can have access to any drug that you need,” says Pat Garcia-Gonzalez, the foundation’s chief executive.

“My next goal,” Garcia-Gonzalez adds, “is to make that possible for all oncology products.” So far, however, the foundation has expanded only into targeted treatments for multiple myeloma and a few types of solid tumour.

### BREAKFAST BENEFITS

Although most of the proposed fixes for the cancer-drug cost conundrum have focused on large-scale systemic change, which often requires buy-in from governments, the pharmaceutical industry and doctor and patient groups, small tweaks also have the potential to make a big difference.

An idea championed by Mark Ratain, director of the Center for Personalized Therapeutics at the University of Chicago in Illinois, is to give expensive cancer drugs with food, rather than on an empty stomach as prescribed. This, he hopes, will improve absorption of the drugs, enabling recipients to lower the dose needed and, therefore, to reduce the cost of treatment.

There are several commonly prescribed cancer pills for which food is known to increase the fraction of the dose that enters the bloodstream, including the lung-cancer drug erlotinib and the melanoma drug vemurafenib. So far, however, Ratain has tested his idea only with the prostate-cancer drug abiraterone.

During abiraterone’s development, trials showed that the concentration and kinetics of the drug differed between people who took it at mealtimes and those who took it without food. The company behind the drug, Janssen Biotech in Raritan, New Jersey, therefore decided to conduct further testing only in the absence of food, to minimize variability between study participants and to reduce the risk of diet-related complications.

The prescribing information for abiraterone, which is marketed as Zytiga, reflects that decision. “Take Zytiga on an empty stomach,” it reads. “Taking Zytiga with food may cause more of the medicine to be absorbed by the body than is needed and this may cause side effects.”

In a pilot study<sup>10</sup> involving 72 people, Ratain and his colleague Russell Szmulewitz, a medical oncologist at the University of Chicago, confirmed this warning by showing that a similar amount of the drug was absorbed when taken as a low dose with a low-fat breakfast as was received with a full dose when fasting. Participants could therefore take one-quarter of the normal dose and still receive the same anti-cancer effects after 12 weeks of treatment,

## IMPROVING CANCER CARE

### The diagnosis differential

In 2015, cancer took the lives of 6 million people in low- and middle-income countries — more than HIV, tuberculosis and malaria combined. Limited access to life-saving medications at an affordable price contributes to this burden. But even if oncologists in these countries could prescribe the same medicines as their better-funded colleagues in the West, it might not reduce the death rate by much.

That’s because, in contrast to residents of wealthier countries, more people with cancer in places such as sub-Saharan Africa seek medical attention only after their tumours have metastasized — the point at which outcomes become poor regardless of the intervention. As David Kerr, a cancer researcher at the University of Oxford, UK, points out: “Unless we can diagnose patients at an earlier stage of presentation, all these ‘fancy-schmancy’ new drugs will have very little impact.”

Early detection is useless, however, if there’s no one who is trained to treat those affected, notes Lawrence Shulman, director of the Center for Global Cancer Medicine at the University of Pennsylvania Abramson Cancer Center in Philadelphia — many developing countries have only a small number of cancer specialists, if any.

To build capacity, since 2011, Shulman has worked with Partners In Health, a non-profit organization based in Boston, Massachusetts, to develop cancer-treatment programmes in Rwanda and Haiti. He has

also engaged in similar work in Botswana.

“You need well-trained nurses. You need appropriate physician expertise,” Shulman says. “You need a certain number of those pieces in place before you can shoot the starting gun and get going.”

A requirement for better training is not limited to countries that are low on resources. The increasing complexity of cancer treatment means that, even in the wealthiest nations, oncologists must work across specialties to achieve optimal outcomes for their patients. That’s why, in 2017, the European Commission’s Expert Group on Cancer Control endorsed the recommendation that all doctors involved in cancer care should undergo a period of cross-disciplinary learning.

Medical oncologists, radiation oncologists and surgical oncologists already collaborate on the day-to-day management of patients through meetings known as tumour boards; such panels of multidisciplinary teams have been shown to increase diagnostic accuracy and to improve patient care.

But Jesper Eriksen, a clinical oncologist at Aarhus University Hospital in Denmark who spearheaded the European recommendation<sup>11</sup>, thinks that there’s still room for improvement. He has called for doctors to complete clinical rotations across disciplines — to enhance the value of meeting other specialists. “Hopefully that will result in a shorter time from diagnosis to treatment,” he says. **ED**

as measured by changes in the level of prostate-specific antigen, a proxy for tumour burden.

If the results hold up to scrutiny, people taking abiraterone will be able to spread the cost of one month’s worth of pills — about \$9,000 — over four months. That could lower US health-care spending by as much as \$20 billion in the next decade, estimates Allen Lichter, former chief executive of the American Society of Clinical Oncology. “The savings that would come simply from taking this with your Cheerios is pretty compelling,” he says.

Last year, with Ratain and others, Lichter co-founded a non-profit organization called the Value in Cancer Care Consortium, which aims to find better and cheaper ways of using existing medicines. The hope is to start by conducting a larger, confirmatory trial of Ratain’s abiraterone study but, according to Lichter, the consortium is struggling to raise the \$5 million needed for a 300-participant trial.

“There’s just a tremendous disconnect at times between what people say is important and what they’re willing to step up to the plate

and make happen,” Lichter notes. “If we can take billions and billions of dollars out of the equation, it cannot help but do good for the cancer patients of the world and for the health-care systems of the world.” Unfortunately, he laments, “Not enough people are focused on value.” ■

**Elie Dolgin** is a science writer in Somerville, Massachusetts.

1. Dusetzina, S. B. *JAMA Oncol.* **2**, 960–961 (2016).
2. Goldstein, D. A. *et al. Oncotarget* **8**, 71548–71555 (2017).
3. Gordon, N., Stemmer, S. M., Greenberg, D. & Goldstein, D. A. *J. Clin. Oncol.* <http://dx.doi.org/10.1200/JCO.2016.72.2124> (2017).
4. Aggarwal, A., Fojo, T., Chamberlain, C., Davis, C. & Sullivan, R. *Ann. Oncol.* **28**, 1738–1750 (2017).
5. Navarria, A. *et al. Value Health* **18**, 131–136 (2015).
6. Garattini, L., Curto, A. & van de Vooren, K. *Eur. J. Health Econ.* **16**, 1–3 (2015).
7. Brennan, H., Kapczynski, A., Monahan, C. H. & Rizvi, Z. *Yale J. Law Tech.* **18**, 275–354 (2017).
8. Shulman, L. N. *et al. J. Clin. Oncol.* **34**, 69–75 (2016).
9. Garcia-Gonzalez, P., Boulbee, P. & Epstein, D. *J. Glob. Oncol.* **1**, 37–45 (2015).
10. Szmulewitz, R. Z. *et al. J. Clin. Oncol.* **35** (suppl. 6S) abstr. 176 (2017).
11. Benstead, K. *et al. Eur. J. Cancer* **83**, 1–8 (2017).